UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL CÓRDOBA

DOCTORADO EN INGENIERÍA
Mención Ingeniería en Sistemas de información

Tesis Doctoral

Estrategia de recomendación por similitud en repositorios con grandes volúmenes de datos de medición y evaluación

Ing. María Laura Sánchez Reynoso

Director: Dr. Mario Diván

Córdoba, Argentina
Año 2023

UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL CÓRDOBA

Comisión de Posgrado
Se presenta esta Tesis en cumplimiento de los requisitos exigidos por la Universidad Tecnológica Nacional para la obtención del grado académico de Doctor en Ingeniería, mención Sistemas de Información.

Estrategia de recomendación por similitud en repositorios con grandes volúmenes de datos de medición y evaluación

Por
Ing. María Laura Sánchez Reynoso

Director: Dr. Mario Diván

Jurados de Tesis
Dr. Horacio P. Leone
Dr. Daniel Riesco
Dr. Javier Darío Orozco

Córdoba, Argentina
Año 2023

A Mario
A mis hijos Ignacio y Santiago, padres y amigos..

Agradecimientos

A lo largo del desarrollo de esta tesis doctoral, he contado con la ayuda de muchas personas a las cuales les debo mi gratitud. Dado que, de no haber sido por ellas, esta tesis no tendría tantos recuerdos, sentimientos y lindas experiencias vividas.

En primer lugar, quiero agradecer a mi director de Tesis Doctoral, el Dr. Mario José Diván, quien además de ser mi compañero de todos los días desde hace 20 años, ha sido una de las personas más importantes en mi formación investigadora y docente. Fue quien desde un principio creyó en mi capacidad y me brindó su apoyo incondicional para que pudiera finalizar este proyecto.

Agradezco a mis hijos, Nacho y Santiago, por ser los pilares fundamentales que siempre están presentes, aconsejándome y animándome constantemente para superar los momentos de incertidumbre.

Agradezco también a mis padres, María del Carmen y Nicolás, por haberme forjado como la persona que soy en la actualidad y por la confianza y seguridad para lograr culminar esta nueva etapa de mi vida.

Agradezco a mis directores de CONICET, Dr. Silvio Gonnet y Dr. Mariano Méndez, por la dedicación y paciencia invertida en la dirección y corrección de diferentes trabajos. Por brindarme las herramientas y los recursos para desarrollar los trabajos necesarios.

Agradezco a mis amigos quienes algunos aún permanecen y otros siguieron sus carreras de vida en sus respectivos lugares, gracias por las compañías, charlas y almuerzos compartidos.

Finalmente, agradezco a la Universidad Tecnológica Nacional – Facultad Regional Córdoba, mi segundo hogar, por el asesoramiento permanente, los cursos y las herramientas recibidas, pero por sobre todo, por la calidez y la calidad humana de las personas que día a día llevan adelante esta noble labor.

¡Muchas Gracias!

Prefacio

La medición es una actividad común a diversas disciplinas, lo que permite determinar el estado actual de un concepto bajo estudio. La evaluación consiste en interpretar los valores medidos a la luz de ciertos criterios para poder concluir (o estimar) sobre su estado. Por tal, el hecho de poder definir proyectos de medición y evaluación es común a diferentes disciplinas, con las particularidades de cada campo de aplicación.

Diversas alternativas se han propuesto para medir en forma remota diferentes conceptos, desarrollando actividades de telemetría sobre ordenadores de placa única (en inglés, Single Board Computers -SBC). El bajo costo y disponibilidad de sensores relacionado con el Internet de las Cosas (IoC) ha permitido masificar la automatización de procesos de medición con una aceptable relación costo-beneficio.

Sin embargo, IoC se caracteriza por la heterogeneidad de las fuentes de datos, y el solo hecho de obtener un valor desde un sensor no aporta demasiado, salvo que exista una lógica embebida en una aplicación o lo interprete una persona con conocimiento en particular.

Por tal, los “framework” de medición y evaluación basados en ontologías permiten incorporar conocimiento específico de los conceptos medidos, sus características, cómo cuantificarlas e interpretarlas. En este sentido, el “framework” C-INCAMI (en inglés Context - Information Need, Concepts, Attributes, Metrics, and Indicators) es un marco conceptual con una ontología subyacente, la cual define los conceptos, términos y relaciones necesarias para especificar un proceso de medición y evaluación (M&E). De este modo y mediante su utilización, se logra un entendimiento común sobre el concepto de lo que representa una entidad, su caracterización a través de atributos, la cuantificación de atributos mediante métricas y la interpretación de los atributos a través de indicadores con su correspondiente criterio de decisión asociados.

La ontología de CINCAMI fue extendida para incorporar la posibilidad de discriminar estados de entidad y escenarios, para extender la definición de los criterios de decisión y su interpretación en función del contexto. De este modo, la estrategia de procesamiento del flujo de datos, es capaz de soportar definiciones de escenarios, análisis de las transiciones en conjunto con la posibilidad de ajustar los criterios de decisión de cada indicador a cada escenario en particular.

Basado en la ontología extendida, se extendió la estrategia GOCAME (Goal - Oriented Context-aware Measurement and Evaluation) derivando en GOCAME -ESVI (Gocame-Entity States, Scenarios, and Visualization Guidelines). La misma permite definir consistentemente un proyecto de medición, especificándolo en un archivo auto-contenido, factible de ser intercambiado entre diferentes SBC para su interpretación y emparejamiento con los sensores. Un componente software denominado adaptador de mediciones (localizado en SBC o dispositivos móviles) es responsable de emparejar la definición de proyectos con los sensores. De este modo, cada sensor tiene asociación con una métrica directa en el proyecto, y por ende sus medidas serán procesadas en consecuencia.

Al momento de procesar e interpretar las medidas de múltiples métricas guiado por sus metadatos (ejemplo, el ID de cada una), un índice de similitud compuesto permite identificar recomendaciones similares, incluso cuando el proyecto no las posea. Basado en la ontología C-INCAMI extendida, se propuso un índice que evalúa la similitud estructural y comportamental de entidades bajo monitoreo y su contexto a partir de la definición del proyecto de medición y su flujo de medidas. Esto permite proveer recomendaciones ante situaciones similares a la actual aun cuando no se cuente con conocimiento específico.

Un esquema de interpretación basado en criterios de decisión soportado por escenario y estados de entidad, es autocontenido en BriefPD. Dado que las fuentes de datos son distribuidas por naturaleza, se planteó una articulación basada en microservicios y arquitectura blockchain para soportar transmisiones indirectas y recomendaciones in-situ.

Complementariamente, se aprovecha la extensión de la ontología de medición para incorporar controles de integridad de datos basados en Merkle Tree a los efectos de poder determinar el origen de los mismos (sin necesidad de que dispositivos de capacidades limitadas deban almacenar los datos transmitidos).

De este modo, se ha logrado integrar en la Arquitectura de Procesamiento basada en Metadatos de Mediciones (en inglés Processing Architecture based on Measurement Metadata - PAbMM), desde la captación de los datos hasta el esquema de recomendación. PAbMM es un motor de procesamiento de flujos de mediciones basada en topologías de Apache Storm, que está especializada en el monitoreo de entidades (físicas o no) en proyectos de medición y evaluación basados en C-INCAMI extendido. Aquí se ha extendido la ontología original (por ejemplo, incorporando escenarios y estados de entidad) y por ende actualizado la arquitectura en consecuencia.

Así, aquí se introduce una estrategia que permite brindar recomendaciones en tiempo real ante situaciones tipificadas basado en los criterios de decisión de los indicadores y en las entidades monitoreadas.

Avances y resultados parciales han sido publicados y validados desde 2018 a la fecha en revistas científicas, capítulos de libros, y conferencias internacionales. A continuación, se listan cronológicamente para su referencia.

- Diván, M. Sánchez-Reynoso, M. Gonnet, S. (2022) **"Measurement Project Interoperability for Real-time Data Gathering Systems**". Future Generation Computer Systems, Elsevier, ISSN 0167-739X, 129, 298-314 https://doi.org/10.1016/j.future.2021.11.031

- Diván, M. Sánchez-Reynoso, (2022). **"Transformations through Blockchain Technology".** Towards a Distributed Record of Measurement Adapters Powered by Blockchain Technology. ISBN 978-3-030-93343-2. pp113-135. Dinamarca. Cham. Springer Cham. https://doi.org/10.1007/978-3-030-93344-9_5

- Diván, M. Sánchez-Reynoso, (2021). **"Big Data Analysis for Green Computing: Concepts and Applications"**. Effect of the measurement on Big Data Analytics. An evolutive perspective with Business Intelligence. ISBN 9781003032328. pp 50-69. Estados Unidos. CRC Press. http://dx.doi.org/10.1201/9781003032328-4

- Diván, M. Sánchez Reynoso, M. Méndez, M. Panebianco J. (2021) "**IoT-based Approaches for Monitoring the Particulate Matter and its Impact on Health"**. e-ISSN: 2327-4662 – IEEE Internet of Things Journal. Institute of Electrical and Electronics Engineers Inc. 2021. Vol. 8, nro 15. pp. 11983 - 12003 http://dx.doi.org/10.1109/JIOT.2021.3068898

- Diván, M. Sánchez Reynoso, M. (2021) "**A Metadata and Z Score-based Load-Shedding Technique in IoT-based Data Collection Systems"**. International Journal of Mathematical, Engineering and Management Sciences. ISSN: 2455-7749. e-ISSN: 2455-7749. Elsevier. Vol.6, nro 1. pp 363 – 382. https://doi.org/10.33889/IJMEMS.2021.6.1.023

- Diván, M. Sánchez Reynoso, M. (2021) **"Metadata-based measurements transmission verified by a Mervle Tree"**. Knowledge-Based Systems. ISSN: 0950-7051. Ed. Elsevier Science BV. Vol. 219. pp 1 -17. http://dx.doi.org/10.1016/j.knosys.2021.106871

- Diván, M. Sánchez Reynoso, M. (2021) **"Strategies based on IoT for supporting the decision making in Agriculture: A Systematic Literature Mapping**". ISSN: 1755-0556 - e-ISSN: 1755-0564 - International Journal of Reasoning-based Intelligent Systems. Vol. 13, nro 3. pp155 – 171. https://dx.doi.org/10.1504/IJRIS.2021.117080

- Diván, M. Sánchez Reynoso, M. Mohd Helmy A., Syed Z., Syed I. (2021) **"IoT based Measurement Collection Distributed Architecture**". Print ISSN: 2516-0281, Online ISSN: 2516-029X. *Annals of Emerging Technologies in Computing (AETiC)*, pp. 1-7, Vol. 5, No. 3, 1st July 2021, Published by International Association of Educators and Researchers (IAER), DOI: 10.33166/AETiC.2021.03.001, http://aetic.theiaer.org/archive/v5/v5n3/p1.html.

- Diván, M. Sánchez-Reynoso, M. Gonnet, S. (2021) **"Recent Applications of Federated Learning in Edge and IoT Environments: A Review**". Proceeding of the 2021 5th International Conference on Information Systems and Computer Networks (ISCON). IEEE. https://doi.org/10.1109/ISCON52037.2021.9702466

- Diván, M & Sánchez Reynoso, M (2020) "**A Real-Time Entity Monitoring based on States and Scenarios**" CLEI Electronic Journal. ISSN 0717-5000. Vol. 23 (1). Pp 2-1:2-25. https://doi.org/10.19153/cleiej.23.1.2

- Diván, M & Sánchez Reynoso, M (2020) "**Optimizing Data Transmission from IoT devices through Weighted Online Data Changing Detectors**" Advances in Data Science and Adaptive Analysis. ISSN 2424-922X. Vol 12 (2). 2041001 (pp.1:33). https://doi.org/10.1142/S2424922X20410016

- Sánchez Reynoso, M & Diván, M (2020) "**Assessment of semantic similarity in entities under monitoring: A systematic literature mapping**". Revista Facultad de Ingeniería Universidad de Antioquía. ISSN 0120-6230. Vol 99. Pp 21-31. https://doi.org/10.17533/udea.redin.20200476

- Diván, M, Sánchez Reynoso, M & Abd Wahab, M (2020) "**Dynamic Switching in the Measurements' Collecting from Heterogeneous Data Sources**". Journal of Physics: Conference Series. ISSN 1742-6596. Vol 1529. 022058:1-8. https://doi.org/10.1088/1742-6596/1529/2/022058

- Diván, M & Sánchez Reynoso, M (2020) "**Relocating the Load-Shedding Strategy in the Data Stream Processing Architecture**" En Congreso Bienal de Argentina (ARGENCON) – IEEE 2020. Resistencia, Chaco. Desde el 1 al 4 de diciembre. https://doi.org/10.1109/ARGENCON49523.2020.9505446

- Sánchez Reynoso, M &Diván, (2020) ***"Applying Data Visualization Guideline on Forest Fires in Argentina".*** 10th International Conference - **CONFLUENCE' 2020**. Department of Computer Science and Engineering. Amity University. Uttar Pradesh, India. January 29 – 31 of 2020. https://doi.org/10.1109/Confluence47617.2020.9058174

- Sánchez Reynoso, M &Diván, (2020) "**Recomendación por similitud semántica en repositorios con grandes volúmenes de datos de medición**". V Jornadas de Intercambio y Difusión de los Resultados de Investigaciones de los Doctorandos de Ingeniería. UTN Córdoba, Córdoba. 6 y 7 octubre. ISBN: 978-950-42-0200-4. https://doi.org/10.33414/ajea.5.751.2020

- Diván, M & Sánchez Reynoso, M (2019) "**An Architecture for the Real-Time Data Stream Monitoring in IoT**". Book Chapter in "Multimedia Big Data Computing for IoT Applications: Concepts, Paradigms, and Solutions", S. Tanwar, S.Tyagi, and N. Kumar (Eds.). pp. 59-100. ISBN 978-981-13-8759-3, Springer. https://doi.org/10.1007/978-981-13-8759-3_3

- Diván, M and Sánchez Reynoso (2019) "**Extending the Data Stream Processing Strategy to Scenario Analysis**". In proceedings of International Conference on Innovations in Computer Science and Engineering (ICICSE).26-28 June of 2019. Miri, Sarawak, Malaysia. International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE). World Academy of Research in Science and Engineering (Publisher). 8(1.4):1-8. ISSN 2278-3091. https://doi.org/10.30534/ijatcse/2019/0181.42019

- Sánchez Reynoso, M & Diván, M (2019) "**Improving the Real-Time Searching in the Organizational Memory**". Procedia Computer Science. Elsevier Ltd. Vol. 154, pp. 293-304. ISSN: 1877-0509. https://doi.org/10.1016/j.procs.2019.06.043

- Diván, M & Sánchez Reynoso (2019) "**A Load-Shedding Technique based on the Measurement Project Definition**". In V. Jain, S. Patnaik, F. Popentiu Vladicescu, and I.K. Sethi (Eds.). Proceedings of 5th International Conference on Intelligent Computing, Communication & Devices (ICCD 2018), Xi'an, China, November 22-24 of 2018. In Advances in Intelligent Systems and Computing, Springer Nature Singapore. pp.1027-1033. ISSN 2194-5357. https://doi.org/10.1007/978-981-13-9406-5_122

- Sánchez Reynoso, M &Diván, (2019) ***"A Systematic Literature Mapping on the Similar Semantically Entities in Measurement Projects".*** The 13th international

conference on e-learning and games" – **Edutainmet 2019.** Pontificia Universidad Javeriana, Cali Colombia. August 15-17 of 2019. http://dx.doi.org/10.1109/ICVRV47840.2019.00033

- Sánchez Reynoso, M &Diván, (2019) ***"Contributions to the Communication of the Official Advertising's Distribution in Argentina".*** 4th International Conference on Information Systems and Computer Networks (**ISCON**). Department of Computer Engineering & Applications, GLA University, Mathura (UP), India. November 21 – 22 of 2019. https://doi.org/10.1109/ISCON47742.2019.9036298

- Diván, M & Sánchez Reynoso, M (2018) "**The Real-Time Measurement and Evaluation as System Reliability Driver**". Book Chapter in "System Reliability Management: Solutions and Technologies". Anand, A & Ram, M (Eds.). CRC Press, Taylor & Francis Group. Pp. 161-188. https://doi.org/10.1201/9781351117661-11

- Diván, M & Sánchez Reynoso, M (2018) "**A library for articulating the measurement streams with columnar data**". International Journal of Engineering and Technology (UAE). Science Publishing Corporation. 7(4.31):234-241. ISSN: 2227-524X. http://dx.doi.org/10.14419/ijet.v7i4.31.23373

Índice de Contenido

Índice de Figuras

Índice de Tablas

Índice de Ecuaciones

Lista de Acrónimos

AM – Adaptador de Mediciones
ASF – Analysis and Smoothing Function
BriefPD – Project Definition
C-INCAMI – Context Information Need, Concepts, Attributes, Metrics, and Indicators
C-INCAMI/MIS – Measurement Interchange Schema
CUSUM – Suma acumulativa
DSE – Data Stream Engines
ECINCAMI – Extended Context Information Need, Concepts, Attributes, Metrics, and Indicators
GF – Función de Recolección de Datos
GOCAME – Goal-oriented Context-aware Measurement and Evaluation
GOCAME-ESVI – Entity, States, Scenarios and Visualization Guidelines
INCITAP – Instituto Nacional de Ciencias de la Tierra y Ambientales de La Pampa
IoC – Internet de las Cosas
IoT – Internet of Things
JSON - JavaScript Object Notation
M&E - Medición y Evaluación
MO – Memoria Organizacional
OMS – Organización Mundial de la Salud
OPA2Vec – Ontologies Plus Annotations to Vectors
OWL – Ontology Web Language
OWL 2 – Ontology Web Language 2
PAbMM – Processing Architecture based on Measurement Metadata
PI – Preguntas de Investigación
PM – Particulate Matter
RQ – Research Questions
RUN – Registro Unificado de Nodos
SBC – Single Board Computers
SMS – Systematic Mapping Studies
SWat – Secure Water Treatment
UNLPam – Universidad Nacional de La Pampa
WSN – Wireless Sensor Networks
XML - eXtensible Markup Language

Capítulo 1	Introducción al Contexto de Estudio

Capítulo 1. Introducción

1.1 Motivación y Antecedentes

Las actividades de medición y evaluación (M&E) son consideradas actividades claves para el conocimiento del estado de situación de una entidad bajo monitoreo. Esto es así tanto en la gestión, como en los procesos industriales, máxime si se pretende construir un modelo que aproxime el patrón de comportamiento de una entidad bajo análisis [1].

La medición y evaluación es un mecanismo a través del cual es posible conocer el comportamiento asociado a una entidad bajo análisis, detectando en forma temprana desvíos y/o anomalías respecto de su conducta esperada [2]. En este sentido, a la hora de implementar el proceso de medición y evaluación, la utilización de marcos formales de M&E, permite acordar los términos, conceptos y relaciones asociadas, a los efectos de promover la consistencia y extensibilidad en dicho proceso [3].

Así, la idea de formalizar el proceso de medición, a través de la utilización de "framework" de M&E, consiste en poder definir conceptos, términos y relaciones entre ellos, que nos permita llevar a cabo su automatización [4].

En el contexto de la automatización del proceso de medición y evaluación, es probable que el origen de los datos sea heterogéneo, tal como sucede en las redes inalámbricas de sensores (en inglés, Wireless Sensor Networks - WSN) [5].Incluso, dichos orígenes tienen la particularidad de que escapan al control del proceso de medición, suelen contar con cierto grado de autonomía e incorporan, en general, capacidades de comunicación entre ellos, tal como ocurre en el Internet de las Cosas (en inglés, Internet of Things - IoT) [6]. Por tal, la homogeneización de los datos provenientes desde fuentes heterogéneas es un aspecto clave para el procesamiento consistente de los mismos (por ejemplo, para computar las temperaturas bajo una unidad y escala homogénea).

El marco de medición y evaluación C-INCAMI (en inglés, Context – Information Need, Concept Model, Attribute, Metric and Indicator) [2], [7] es un marco conceptual con una ontología subyacente, la cual define los conceptos, términos y relaciones necesarias para especificar un proceso de M&E. De este modo y mediante su utilización, se logra un entendimiento común sobre el concepto de lo que representa una entidad, su caracterización a través de atributos, la cuantificación de atributos mediante métricas y la interpretación de los atributos, a través de indicadores con su correspondiente criterio de decisión asociados. Por ende, cada proyecto de medición y evaluación puede ser definido en términos del marco C-INCAMI, empleando como guía la estrategia GOCAME (Goal-Oriented Context-Aware Measurement and Evaluation)[8]. La estrategia GOCAME es una estrategia multipropósito, orientada al contexto, que establece las actividades y fases necesarias para definir un proyecto de medición y evaluación en términos del marco C-CINCAMI.

El esquema de intercambio de mediciones basado en el marco C-INCAMI, se conoce como C-INCAMI/MIS (en inglés, C-INCAMI/Measurment Interchange Schema) [9], y permite intercambiar conjuntamente los datos de la medición y los datos descriptivos del proyecto de M&E. Ello posibilita embeber metadatos de la medición junto con las medidas, lo que beneficia la detección de inconsistencias mediante contraste con la definición del proyecto. Por ejemplo, si a través de la medición vinculada con una métrica de humedad ambiente, se informare desde el sensor un valor de 1000, con la sola utilización de la definición de la métrica (esto es conociendo su escala, unidad, método de medición, entre otros aspectos), sería posible detectar la anomalía inmediatamente sobre su recepción (sea que se haya producido por error de calibración, o bien, por interferencia en el ambiente).

Los motores de procesamiento de flujo de datos (en inglés, Data Stream Engines - DSE) procesan una secuencia de datos en tiempo real a medida que estos arriban [10]. En este contexto, los recursos disponibles para el procesamiento de datos son limitados, lo que implica que cada dato deba ser leído hasta una vez en lo posible, ya que una segunda lectura sería contraproducente debido a que un dato más actualizado estaría esperando para ser procesado. De este modo, los orígenes de datos asociados con los DSE se caracterizan por ser ilimitados, impredecibles, independientes del DSE, con tasa de arribo eventualmente volátil y donde los datos son susceptibles de incorporar errores [11].

La arquitectura de procesamiento basada en metadatos de medición (en inglés, Processing Architecture Based on Meassuremet Metadata - PAbMM) [12], es un motor de procesamiento de flujo de datos en tiempo real, especializado en proyectos de evaluación y medición. Dichos proyectos son definidos empleando la estrategia GOCAME juntamente con el marco C-NCAMI. Los flujos de datos a procesar están organizados bajo el esquema C-INCAMI/MIS, lo que promueve su consistencia y comparabilidad, como así también la incorporación de fuentes de datos heterogéneas.

La memoria organizacional (MO) es un repositorio de grandes volúmenes de datos, empleado por PAbMM [13] para: a) almacenar la definición de los proyectos de M&E en base a C-INCAMI, b) documentar la experiencia previa y el conocimiento de los expertos, en relación a los criterios de decisión vinculados con los indicadores y los cursos de acción asociados, y c) almacenar el volumen histórico de las mediciones cuando sea requerido (lo que dará origen a un repositorio con grandes volúmenes de datos provenientes de los flujos de datos asociados). Ello permite la incorporación de razonamiento basado en casos sobre la memoria organizacional, para detectar situaciones similares y recomendar cursos de acción ante la ocurrencia/configuración de una situación dada.

El flujo continuo de mediciones procesado por PAbMM, permite retroalimentar la MO cuando sucedan situaciones no tipificadas [13]. Ello es útil, ya que PAbMM emplea la experiencia previa y el conocimiento de expertos desde la MO en memoria, para

recomendar cursos de acción cuando las situaciones tipificadas sucedan en una entidad bajo monitoreo dada. A su vez, las no tipificadas permiten incorporar en la MO situaciones no contempladas a priori, y que eventualmente podrían ser de interés para los expertos en relación con la entidad bajo monitoreo.

Ahora bien, el contexto de búsqueda sobre un repositorio con grandes volúmenes datos es completamente diferente del vinculado con los DSE [10], [14] [15]. Es decir, en un repositorio con grandes volúmenes de datos basado en almacenamiento persistente, el rendimiento de cada medio de almacenamiento podría ser heterogéneo, los tiempos de acceso (lectura/escritura) son completamente diferentes del acceso en memoria y se cuenta con una mayor la tolerancia respecto del tiempo de respuesta ante una consulta dada. En los motores de procesamiento de flujo de datos, el dato será leído y procesado hasta una vez, las decisiones se toman en tiempo real, los tiempos de acceso están asociados con la memoria y todo ello deberá ocurrir dentro de los límites impuestos por los recursos disponibles en ese instante (por ejemplo, memoria, procesador, etc.)
De este modo, la memoria organizacional constituye un contexto diferente de procesamiento respecto de PAbMM. Motivo por el cual es necesario un mecanismo de búsqueda en memoria, a partir del conocimiento de expertos y la experiencia previa de cada proyecto, para poder capitalizar y aplicar la recomendación de cursos de acción como soporte al proceso de toma de decisiones en tiempo real de PAbMM. Ahora bien, cuando una entidad no posee eventualmente experiencia previa, la idea en PAbMM [16] consiste en reducir el espacio de búsqueda en memoria respecto de las experiencias previas y conocimiento de las entidades vinculadas. Así, aun cuando no exista experiencia previa específica, PAbMM, podrá recomendar curso de acción vinculados con entidades cuya estructura y comportamiento son similares a la monitoreada y recomendar por analogía.

El análisis de similitud estructural se calcula a partir de los metadatos asociados con la definición del proyecto de M&E [17] basado en C-INCAMI. Sin embargo, puede suceder que dos entidades estructuralmente similares, se comporten en forma diferente. En tal caso, dicha situación es abordada mediante el coeficiente de similitud comportamental, el cual es obtenido mediante análisis de correlación multivariado a partir de las métricas que cuantifican el comportamiento de la entidad [16]–[18].

Teniendo en cuenta lo antes mencionado, en lo que se refiere a los coeficientes (estructurales y comportamentales) y a las entidades con y sin experiencia previa, la pregunta a establecer es ¿De qué modo es posible detectar entidades semánticamente similares en un proyecto de M&E basado en C-INCAMI, aun cuando presenten diferencias aparentemente estructurales? Es decir, a partir de la definición de las características de cada entidad ¿Cómo es posible determinar si refieren aproximadamente al mismo concepto? Si fueren semánticamente similares (no necesariamente equivalentes) ¿Existe experiencia previa que permita utilizar el conocimiento previo de una en la otra? De este modo, el objetivo principal de la

presente tesis consiste en desarrollar una estrategia de recomendación en memoria, a partir de repositorios de medición y evaluación basados en el marco formal C-INCAMI, a los efectos de localizar entidades bajo monitoreo semánticamente similares, y poder así, reutilizar su conocimiento y experiencia en el proceso de toma de decisiones en tiempo real cuando sea requerido.

1.2 Planteamiento del Problema

La memoria organizacional en PAbMM incorpora a) la definición del proyecto de medición y evaluación en términos del marco C-INCAMI, b) el conjunto de medidas históricas para una entidad dada, y c) la experiencia previa y el conocimiento específico para cada entidad bajo monitoreo. La memoria organizacional emplea la experiencia previa y-o conocimiento específico sobre una entidad dada, para brindar recomendaciones cuando alguna de las situaciones tipificadas sucede.

Ahora bien, cuando una entidad presenta una nueva situación la cual no posee antecedentes asociados, o bien, la entidad es nueva y carece de historia, la recomendación en PAbMM es guiada localizando entidades estructuralmente y comportamentalmente similares. El problema es que estas dos últimas variantes para localizar las entidades similares (y por ende su experiencia/conocimiento asociado), suponen que la definición de su estructura y el comportamiento asociado es homogénea semánticamente.

Justamente el punto es que este último aspecto es un supuesto de difícil cumplimiento, ya que los proyectos de M&E bajo monitoreo pueden ser variados, al igual que sus entidades e incluso podrían emplear sinónimos en la definición de atributos. De este modo, es poco probable que en dos proyectos que monitorean entidades, se tengan dos atributos que sean definidos narrativamente en español exactamente en los mismos términos y sin variante (en tal caso, sería duplicación). Lo que sí sería probable, es que en dos proyectos dados se definan dos atributos de entidades diferentes empleando conceptos eventualmente equivalentes (o mediante sinónimos) para los atributos. Por ejemplo, es posible que se defina sintácticamente un atributo como “Temperatura corporal” para una entidad “paciente trasplantado ambulatorio” el cual es monitoreado por el proyecto A y B. En el proyecto A define la temperatura corporal axilar, pero en el proyecto B define la temperatura corporal sublingual. Esta situación estructuralmente y comportamentalmente serían correctas, pero semánticamente son diferentes. Es decir, el parámetro para determinar si un paciente tiene “fiebre” podría cambiar, y con ello los criterios de decisión e indicadores del tomador de decisión. Así, se está frente a un problema de conceptos asociados con la definición narrativa en español de cada atributo que caracteriza a la entidad bajo análisis dentro del proyecto de M&E.

La otra parte del problema es que incluso determinadas las entidades similares semánticamente, la experiencia y-o conocimiento previo debe estar organizado y

sintetizado “en memoria” (no en una base de datos columnar persistente) a los efectos de servir tales recomendaciones en tiempo real al tomador de decisiones dentro de PAbMM.

Resumiendo hasta aquí, el problema es la imposibilidad de determinar la similitud semántica a partir de los atributos que caracterizan narrativamente en español las entidades bajo monitoreo en proyectos de M&E basados en C-INCAMI, y a partir de ello, el inconveniente en organizar y sintetizar tal experiencia previa y/o conocimiento por similitud semántica en memoria de forma tal, que permita servir como recomendación aproximada al tomador de decisiones en tiempo real cuando no existe referencia previa sobre una entidad.

Esta situación lleva a limitaciones de precisión en las recomendaciones que brinda actualmente PAbMM, la duplicación de atributos y/o categorías de entidad en la definición de proyectos, el incremento de la complejidad de proyectos a partir de la reiteración de conceptos, la imposibilidad de detectar conceptos contradictorios en las definiciones de atributos, y las limitaciones en acotar el espacio de búsqueda basado en el emparejamiento de conceptos.

Esto conlleva a trabajar sobre alguna estrategia de análisis semántico sobre la definición de los atributos que caracterizan la entidad bajo monitoreo, para determinar si son similares o no en concepto, por ejemplo, detectar sinónimos.

La posibilidad de identificar sobre la definición de proyectos de M&E basados en C-INCAMI cuando las entidades son semánticamente similares, permitiría desarrollar una estrategia de recomendación en memoria que limite el espacio de búsqueda basado en la similitud semántica, lo cual permitiría focalizar en experiencia/contenido específico.

Claro que, en este último caso, la búsqueda en otro medio diferente a la memoria sería inviable para PAbMM dado que este procesa en tiempo real. De este modo, el análisis de similitud semántica debiera permitir la organización y síntesis del contenido en memoria para brindar la recomendación sin necesidad de acceder al repositorio persistente.

Mucho se ha trabajado en términos de búsqueda por similitud, incluso en el ámbito de los motores de búsqueda dentro del área de recuperación de la información, lo cual es un área en constante evolución. En este caso en particular, los proyectos de M&E están definidos sobre una estructura basada en la ontología de C-INCAMI. Sobre ella, debiera procesarse el lenguaje natural en español para determinar si la definición narrativa en español de un atributo dado es similar o no a otro, y a partir de allí, arribar a un scoring de similitud semántica entre entidades que luego permitirá generar el esquema de organización y síntesis en memoria para las recomendaciones que soportarán al tomador de decisiones de PAbMM.

Los supuestos sobre los cuales se organiza el presente trabajo de tesis son:

- Todos los proyectos son específicos de medición y evaluación

- Los proyectos de medición y evaluación se definen en términos del marco C-INCAMI
- La definición de los atributos que caracterizan la entidad bajo análisis:
  - Se encuentra en español,
  - Está acotada a una cantidad de caracteres dada,
  - No contiene abreviaturas
  - Puede referir a la definición de otros conceptos de uno u otro proyecto de M&E.
- Los datos llegan a la arquitectura de procesamiento:
  - Organizados según el esquema C-INCAMI/MIS,
  - Desde fuentes de datos posiblemente heterogéneas de las cuales no se tiene control,
  - Sin cota o límite de volumen de medidas
- Se emplea PAbMM como arquitectura de procesamiento a los efectos del desarrollo del análisis de similaridad, procesamiento e implementación de las estructuras organizativas y búsqueda en memoria,
- Las decisiones se toman en tiempo real y las recomendaciones son adjuntadas para sugerir cursos de acciones posibles.

1.3 Hipótesis

Con este trabajo de tesis, se persigue establecer una estrategia que permita buscar en memoria entidades bajo monitoreo similares semánticamente, y a partir de ellas, organizar y sintetizar las recomendaciones en memoria para brindar una respuesta aproximada al tomador de decisiones cuando no existe conocimiento/experiencia previa en una entidad. Es un tema que ofrece numerosas posibilidades de investigación y desarrollo, y dentro de esas posibilidades, el presente trabajo orienta su investigación teniendo en cuenta las siguientes interrogantes:

- ¿Es posible generar un scoring de similitud semántica a partir de entidades bajo monitoreo cuyos atributos se definen narrativamente en español en proyectos de M&E basados en C-INCAMI?
- A partir de las entidades semánticamente similares ¿Sería posible organizar y sintetizar el conocimiento/experiencia previa en memoria para mejorar la precisión de las recomendaciones en PAbMM?

En este sentido, el presente trabajo de tesis postula que:

1.) Si a partir de proyectos de M&E basados en C-INCAMI es posible determinar un scoring de similitud semántica entre las entidades bajo monitoreo,

2.) y a partir de ello, incorporar una organización y síntesis específica de la experiencia previa/conocimiento limitado a la memoria disponible,

entonces se podrá identificar entidades similares semánticamente, lo que permitiría incrementar la precisión de las recomendaciones, responder con recomendaciones en forma aproximada cuando no se cuente con conocimiento/experiencia previa, detectar definiciones contradictorias en los proyectos, y acotar el espacio de búsqueda de la memoria organizacional persistente (cuando sea requerido offline) solo a aquellas entidades semánticamente similares.

1.4 Objetivo del trabajo de tesis

1.4.1 Objetivo General

Desarrollar una estrategia de recomendación en memoria, basado en entidades bajo monitoreo semánticamente similares, para mejorar la precisión y reutilización de conocimiento y/o experiencia previa ante situaciones nuevas y-o no tipificadas, en los cuales una decisión requiera de cursos de acción como soporte.

1.4.2 Objetivos Específicos

En lo que respecta a los objetivos específicos, se enumeran los siguientes a continuación:

- Definir la estrategia de análisis de Similitud Semántica en entidades asociadas con proyectos de M&E basados en C-INCAMI
- Implementar el Scoring Semántico a partir de la estrategia de análisis de similitud semántica
- Definir la estrategia de priorización de contenido (conocimiento/experiencia previa) en base al scoring semántico de entidades bajo análisis,
- Diseñar los mecanismos de organización y síntesis en memoria de los contenidos para recomendación
- Implementar la estrategia de recomendación aproximada basada en similitud semántica ante el tomador de decisiones, cuando una situación no tipificada o sin antecedentes se presenta.

1.5 Principales Contribuciones

A partir del Desarrollo de una estrategia de recomendación en memoria, basado en entidades bajo monitoreo semánticamente similares, se pretende:

1. Mejorar la precisión en las recomendaciones ante el tomador de situaciones *para una situación dada,*

2. *Reutilizar el conocimiento y/o experiencia previa ante situaciones nuevas y-o no tipificadas, basado en similitud semántica de las entidades bajo monitoreo,*
3. *Detectar definiciones contradictorias que pudieren afectar las recomendaciones asociadas a una decisión,*
4. *Detectar atributos homónimos y evitar la redundancia entre atributos en sus definiciones,*
5. *Acotar el espacio de búsqueda en el repositorio columnar con grandes volúmenes de datos a partir de la similitud semántica de entidades bajo monitoreo,*

Como resultado de la investigación y el trabajo a realizar, se pretende incorporar en PAbMM la estrategia de recomendación *basada en similitud semántica de entidades bajo monitoreo, orientada a mejorar la precisión en las recomendaciones aproximadas ante el tomador de decisiones.*

1.6 Estructura de la Tesis

La estructura de los capítulos crea la tesis lógicamente del siguiente modo:

- **Capítulo 2: Estado del Arte**

  Se aborda el Estado del Arte donde inicialmente se introduce al concepto de sistemas de recomendación, se incursiona en marcos de medición y evaluación, se hace referencia a las medidas de similitud como así también a conceptos referidos a la memoria organizacional.

- **Capítulo 3: Formalización del proyecto de medición y evaluación**

  Se desarrolla el capítulo en cuatro secciones donde, en primer lugar se discuten los conceptos, términos y relaciones entre los mismos considerando un abordaje ontológico. Seguidamente se discuten nuevos conceptos que fueron incorporados al marco de medición original. Se introduce en la segunda sección una evolución de la estrategia GOCAME con el objetivo de dar soporte a nuevos conceptos ontológicos e incorporar guías de visualización para los indicadores. Un nuevo esquema de organización y formato de proyectos de medición y evaluación es introducido y finalmente se plantea la integración del marco de medición extendido, la estrategia y el nuevo formato del proyecto de medición y evaluación en el proceso de recolección de datos.

- **Capítulo 4**: **Medidas de similitud para proyectos de medición y evaluación**

  En este capítulo la idea central consiste en cómo estimar la similitud entre proyectos de medición y a partir de dicha similitud poder ordenar en forma descendente los proyectos que sean similares. Se organiza el capítulo en cinco secciones, donde en la primer sección se describe la similitud estructural. En la segunda sección se lleva a cabo un análisis de la similitud comportamental. La tercera sección se centra en realizar una comparación entra las similitudes estructurales y comportamentales. En la sección cuarta, se analizan los resultados obtenidos de una simulación discreta para finalmente en la sección quinta esbozar las conclusiones correspondientes al capítulo.

- **Capítulo 5**: **Arquitectura de Procesamiento (PABMM)**

  Se describe en este capítulo la arquitectura del procesamiento de datos basado en metadatos de medición para obtener una visión global del mismo. También se analizarán los detectores de cambios incrementales, se detallarán los modos en que los escenarios y estados de entidad se calculan en tiempo real como así también se abordarán las funcionalidades por capa y nivel junto con las estrategias de búsqueda. Se organiza el capítulo en siete secciones, donde en la primer sección se describen términos generales de la arquitectura. En la segunda sección, se aborda la recolección de datos y su vinculación con al definición del proyecto. La tercera sección hace mención al procedimiento de determinación de estados y escenarios. En la cuarta sección, se aborda el proceso de flujos de medidas. La sección quinta describe el análisis realizado de las medidas obtenidas. En la sección sexta se introduce el rol de la toma de decisiones y recomendaciones para finalmente en la séptima sección describir las conclusiones asociadas al capítulo.

- **Capítulo 6: Tecnologías de Soporte a la Arquitectura de Procesamiento**

  En el presente capítulo se introducen los detectores de cambio en el adaptador de medición. Se describe el descarte selectivo y se plantean organizaciones de búfer particulares como así también el uso de árboles de Merkle. Se organiza el capítulo en cinco secciones, en donde en la primer sección se desarrolla conceptos asociados con transmisión de datos y detectores de cambio. En la

segunda sección, se describe el descarte selectivo y los metadatos de medición. La tercera sección plantea el registro de integridad basado en el árbol de Merkle. En la cuarta sección se introduce al registro distribuido del adaptador de mediciones para finalmente indicar en la sexta sección, las conclusiones generales del capítulo.

- **Capitulo 7: Escenario de uso - Monitoreo de Material Particulado**

  En este capítulo se describe un escenario de uso para la arquitectura de procesamiento de datos, haciendo foco en material particulado. El capítulo se encuentra organizado en cinco secciones, donde en la primer sección, se introduce el concepto de material particulado, su impacto e importancia. La segunda sección, describe los dispositivos y la estrategia para realizar el monitoreo del material particulado. En la tercera sección se describe el escenario de uso en la provincia de La Pampa. La cuarta sección desarrolla la aplicación de la distancia compuesta para finalmente en la sección quinta plantear las conclusiones al respecto.

- **Capítulo 8:** Sintetiza las principales contribuciones efectuadas en la presente tesis y finaliza con los lineamientos fundamentales de los potenciales trabajos futuros que se desprenden de la misma.

Capítulo 2	Estado Del Arte

Capítulo 2. Estado del Arte

Introducción

En el desarrollo del presente capítulo, se consideró aplicar para el análisis, la metodología del Estudio de Mapeo Sistemático (en inglés Systematic Mapping Studies - SMS) [19]. El objetivo es realizar un mapeo sistemático de la literatura para identificar y encontrar evidencia de los conceptos a desarrollar en el presente capítulo.

El Estudio de Mapeo Sistemático (SMS) establece una serie de guías [20]–[23] para guiar la investigación, orientado a buscar toda la evidencia en un dominio amplio y en este sentido se aplica a marcos de medición, procesamiento de flujo de datos y a la ciencia de los datos.

El protocolo a utilizar para llevar a cabo la revisión sistemática consiste en los siguientes pasos: a) Identificar una necesidad de revisión, b) Especificar las preguntas de investigación, c) Determinar la estrategia de búsqueda, d) Especificar el proceso de extracción de datos, e) Especificar el proceso de síntesis y f) Revisar el desarrollo del protocolo. El objetivo al realizar el mapeo sistemático consiste en poder identificar métodos y técnicas asociadas con los conceptos de sistemas de recomendación, marcos de medición y evaluación, medidas de similitud y memoria organizacional. De acuerdo con la metodología se deben definir las preguntas de investigación (en inglés Research Questions - RQ) que guiarán el trabajo. Seguidamente se debe definir la cadena de búsqueda y luego describir la estrategia de búsqueda junto con los parámetros asociados para realizarla. En este punto, cabe mencionar que la base de datos seleccionada para llevar a cabo la estrategia de búsqueda es Scopus, centrándose en artículos, documentos de revistas y capítulos de libros. El proceso de selección de los artículos obtenidos a partir de la ejecución de la cadena de búsqueda, tiene como puntos clave lo siguiente: a) Eliminar los artículos duplicados, b) Filtrar los artículos considerando criterios de inclusión y exclusión y c) Confeccionar una lista con los artículos resultantes luego de la lectura de los mismos.

Descripta la metodología a utilizar, el presente capítulo abordará el estado del arte en donde se describirá el contexto de trabajo de la tesis, considerando aspectos asociados a los sistemas de recomendación, los cuales se describen en el apartado 2.1. Por otro lado, se llevará a cabo un análisis de los diferentes marcos de medición y evaluación encontrados, permitiendo determinar las ventajas de cada uno de ellos en comparación con el marco a utilizar para el desarrollo de la presente tesis.

Por su parte en la sección 2.3, se introduce a los conceptos de medición, similitud, semántica, entre otros. En tanto en la sección 2.4 se abordará los aspectos importantes asociados a la memoria organizacional que serán considerados a los efectos del presente trabajo. Finalmente se plantearán conclusiones a las cuales se arribaron en el capítulo.

2.1 Sistema de Recomendación

En la actualidad, mucho se habla de sistemas de recomendación, pero es importante conocer cuál es el concepto de los mismos, cómo funcionan y los tipos que existen. Estos sistemas tienen su origen a mediados de la década del ´90 y han ido tomando especial relevancia en estos tiempos dado que permiten entre tantas cosas, incrementar la venta de productos, incrementar la satisfacción de los clientes brindando recomendaciones de acuerdo al perfil de los diferentes usuarios [24] [25].

De acuerdo a [26] se define como sistema de recomendación a las técnicas y herramientas que establece un conjunto de criterios sobre los datos de los usuarios para poder realizar sugerencias de elementos que para el usuario sea de utilidad o de valor. Dichas sugerencias se encuentran relacionadas con el proceso de toma de decisiones, ejemplo, qué noticias leer o qué producto comprar [24]. Dichos sistemas seleccionan los datos proporcionados para analizarlos y luego transformar los mismos en conocimiento de recomendación. La importancia de estos sistemas hoy en día, se debe a que las recomendaciones que brindan surgen de asociar los perfiles de los usuarios a través de un historial de compras o bien a través de la selección de determinados contenidos, obteniendo como resultado un alto grado de eficiencia en cuanto a los gustos, preferencias o necesidades de los usuarios. Anteriormente las plataformas de contenido como las de venta de productos, funcionaban utilizando rankings, pero no era posible personalizar la experiencia del usuario a través de sus intereses, de allí que dichos sistemas de recomendación han evolucionado considerando las experiencias, intereses y perfiles de los usuarios.

A partir de dicha evolución, se establecieron diferentes tipos de sistemas de recomendación [27], los cuales dependen de ciertos factores y variables que forman parte de su funcionamiento. Entre los cuales encontramos los siguientes: **a) Sistemas basados en datos:** estos sistemas se suelen implementar en por ejemplo, ventas de productos y toman como referencia la popularidad del objeto de estudio, por ejemplo el número de ventas. Las recomendaciones en este tipo de sistemas, se basan en el conocimiento anterior, el usuario es quien determina qué es lo que desea, por lo tanto el sistema basándose en su base de datos, determina aquellos ítems que pudieran satisfacer los requerimientos del usuario. Tienen como ventaja que se los considera fáciles de implementar y presentan un cierto grado de efectividad. Sin embargo, la desventaja que presentan es la incapacidad de personalizar los criterios para recomendar al usuario de acuerdo a sus preferencias. **b) Sistemas basados en contenido:** son aquellos que toman ciertos datos del historial del usuario e intenta predecir la búsqueda del usuario y realiza algunas sugerencias similares para el mismo, de acuerdo al contenido registrado. Es decir, que las recomendaciones que proporciona están basadas en la información del sistema y no en la del usuario real, son muy predictivos. Estos sistemas suelen armar un perfil considerando los intereses de los usuarios por lo que las recomendaciones o sugerencias que brinda, están concentradas en las características de los ítems valorados. Por ejemplo, si un usuario reproduce

música de un determinado artista, en su perfil se guardarán las canciones asociadas a dicha preferencia, provocando esto que al momento de recomendarle nuevas canciones, el artista de su preferencia tendrá prioridad en la lista de canciones a reproducir. En la actualidad estos sistemas son los que más presencia tienen, dado que una de las ventajas que presentan es que permiten descubrir opciones que se adapten a determinadas características de productos o contenido y de esta manera elegir opciones que sean similares. **c) Sistemas basados en colaboración:** este tipo de sistemas son novedosos dado que genera recomendaciones a través del análisis de los datos lo que permite contrastar la información del perfil del usuario y la de un conjunto de usuarios. Recomiendan a un usuario activo basándose en las opiniones de otros usuarios. Estos tipos de sistemas, utilizan algoritmos como el de vecino más cercano [28], donde los datos recolectados permiten estimar similitudes entre usuario o grupo de ellos, y luego efectuar la recomendación de acuerdo a los criterios establecidos.

La ventaja de esto, es que permite al sistema agrupar perfiles similares y aprender de los datos que recibe en forma general, con ello se pueden brindar recomendaciones individuales.

A continuación se puede observar en la Tabla 1, las principales características que presentan cada uno de los diferentes tipos de sistemas de recomendación mencionados previamente.

Tabla 1. Comparación entre las principales características de los sistema de recomendación

Tipo de Sistema	Basados en datos	Basado en contenido	Basado en colaboración
Características Principales	Análisis de datos	Perfil del historial del usuario	Agrupa perfiles similares de usuarios
	Fácil de implementar	Predice la búsqueda del usuario	Aprende de los datos generales
	Recomendaciones sin personalizar	Sugerencias similares	Recomendación individual

Estos sistemas de recomendación trabajan a través de la aplicación de diferentes técnicas, las cuales serán aplicadas de acuerdo al contexto de conocimiento donde se utilizarán dichos sistemas. Esto quiere decir que dependiendo del problema a solucionar, se aplicará una determinada técnica que presente ventajas ante la situación analizada.

Los sistemas de recomendación o sistemas de recomendación, proveen a los usuarios una lista de sugerencias de acuerdo a sus preferencias que pueden gustarles. Sin embargo, la mayoría de estos sistemas son vulnerables y presentan algunas limitaciones vinculadas con la recomendación [29]. La idea es que el papel de dicho sistemas sea definir una serie de recomendaciones orientadas en forma general a las preferencias de

los usuarios permitiéndoles de este modo descubrir contenidos atractivos y disminuir el tiempo de búsqueda de los mismos.

Por otro lado, cuando se habla de sistemas de recomendación, es importante tener en cuenta que un proceso de recomendación, presenta ciertas fases, tal como se muestra en la Figura 1:

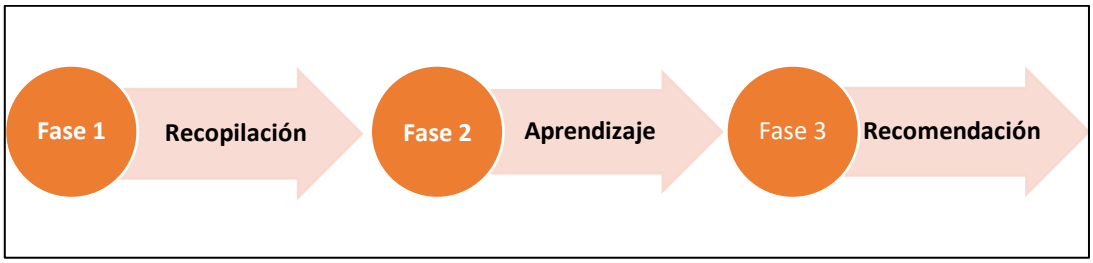

Figura 1. Fases de un proceso de recomendación

1. **Recopilación de datos:** esta primer fase consiste en recopilar los datos provenientes de calificaciones, comentarios sobre productos, historial de búsqueda, entre otros.

2. **Aprendizaje:** durante esta fase se aplica un algoritmo de aprendizaje el cual permite filtrar las funciones del usuario a partir de los datos recopilados.

3. **Fase de Recomendación:** finalmente en esta última fase, el objetivo principal es realizar la recomendación a partir de los datos asociados a las preferencias de los usuarios.

Una vez descriptas las fases correspondientes al procesos de recomendación, se puede observar en la Figura 2 la descripción del esquema general asociado a dicho proceso:

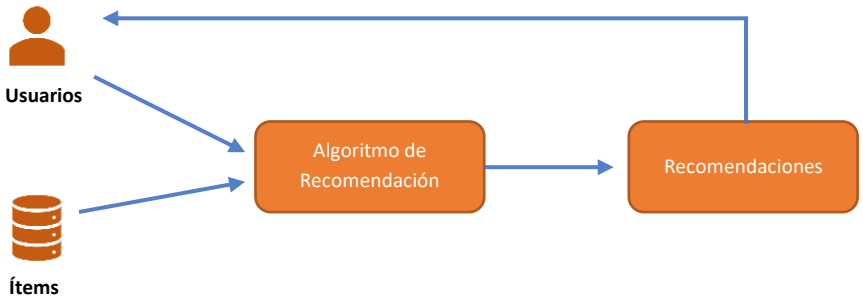

Figura 2 Proceso de Recomendación - Esquema General

En la figura anterior, podemos observar que los ítems o datos de entrada son proporcionados por el usuario quien con dicha información que suministra, da comienzo al proceso de recomendación. El paso siguiente consiste en combinar los datos a través de la utilización de algoritmos para que el modelo de recomendación genere las sugerencias al usuario de acuerdo al perfil del mismo.

El interés por los sistemas de recomendación, ha ido incrementándose en el último tiempo, a tal punto que las empresas no conciben sus negocios sin ellos. Así podemos mencionar Amazon.com donde se recomienda al usuario por ejemplo libros y todo tipo de productos de acuerdo a sus preferencias [30]. Por otro lado, la plataforma Netflix en la sección películas, presenta uno de los sistemas de recomendación con mayor incidencia, donde considera aspectos tales como qué títulos agregó a su lista de usuario, cuáles fueron las películas que vio y le gustó. Por último, la plataforma de música Spotify, realiza recomendaciones personalizadas a sus usuarios considerando aspectos como ser qué canciones se encuentra escuchando o bien cuáles se encuentran en su lista de reproducción, canciones repetidas, hora del día, idiomas, géneros entre otros, que permiten ofrecer recomendaciones e ir mejorando en forma gradual la lista de artistas y canciones que pudieran gustarle al usuario a medida que se avanza en el análisis de dichos perfiles musicales.

Las plataformas mencionadas previamente, cuentan con sistemas de recomendación los cuales utilizan técnicas, métodos y modelos que permiten obtener buenos resultados. Sin embargo, pueden presentar algunas falencias a la hora de implementarlos [31], tales como: a) incorporar en forma constante datos asociados a los ítems, usuarios y valoraciones, lo que provoca un crecimiento de los mismos; b) recomendar a un usuario nuevo de quien no se cuenta con datos que permitan sugerir de acuerdo a sus preferencias; y c) clasificar las preferencias del usuario para poder brindar un recomendación adecuada teniendo en cuenta su perfil.

Previamente se realizó una introducción general a conceptos asociados a los sistemas de recomendación, como ser clasificación, características y formas de trabajo de los mismos. Sin embargo, a los efectos de la presente tesis, es necesario abordar conceptos relacionados con Internet de las Cosas (IoC) y su vinculación con nuestro trabajo. Por ello, es importante mencionar que IoC describe un sistema complejo. Se asocia un sistema de componentes interrelacionados y conectados que permite abordar diferentes escenarios y ofrecer soluciones coherentes en muchos ámbitos (por ejemplo gubernamental, comercial, público industrias privadas, etc.) [32]. El IoC se propuso a principios de la década 1990, cuando el británico Kevin Ashton (Un ejecutivo de Procer & Gamble) tomó la iniciativa de formar un grupo de investigadores que se dedicaban a buscar información sobre el sistema de identificación por radiofrecuencia (RFID) y las tecnologías de sensores [33]. De este modo, la IoC podría definirse como un conjunto de objetos conectados permanentemente a un sistema inteligente y digital, cuya principal tarea es gestionar grandes volúmenes de información. Existen varias definiciones asociadas al IoC. Cada una de ellas ofrece diferentes perspectivas en función del área de aplicación y de la percepción sobre el concepto subyacente. Así, IoC se asocia a un sistema complejo en el que los dispositivos, sensores u objetos están interconectados, permitiendo almacenar información que posteriormente será necesario transmitir a otros componente de la red [34]. En la actualidad, cuando se habla de IoC, se considera que aún está en su fase inicial. Sin embargo, su uso está avanzando

rápidamente , y cada vez es más común observar los diferentes campos de aplicación que tiene. Es posible encontrar diferentes aplicaciones de IoC, por ejemplo 1) Monitorización de pacientes: los médicos recogen datos de los pacientes a través de sensores. Permite a los médicos estar informados sobre las condiciones de los pacientes en tiempo real. Su objetivo es implementar un esquema de alertas para prevenir eventos letales en los pacientes [35]; 2) Smart Farming: en la agricultura, los sensores IoC permiten obtener información sobre las condiciones ambientales (por ejemplo, suelo, humedad, temperatura, etc.). Ayuda a los agricultores a controlar el riego, determinar el mejor momento para sembrar, así como descubrir si hay enfermedades en el suelo o en las plantas [36], [37]; 3) También es posible de encontrar una gran cantidad de tecnologías basadas en el IoC aplicadas a la detección de incendios forestales, la prevención de inundaciones, la gestión de la seguridad, el transporte, la evaluación del grado de contaminación del aire en las grandes ciudades, entre otras [38].

El IoC ha surgido como una alternativa segura, barata y de bajo coste para aumentar el área de cobertura y la resolución en muchas aplicaciones de monitorización. Las necesidades de monitorización se han abordado mediante la articulación de dispositivos basados en el IoC junto con procesos de medición bien definidos que conducen a la implementación de diferentes sistemas de recogida de datos en tiempo real [39]. Permite que mediante el análisis de los datos recibidos, se pueden emitir recomendaciones en función de la situación detectada, proporcionando la posibilidad de prevenir y detectar cualquier anomalía que pueda surgir del análisis de datos durante el proceso de medición.

Los recopiladores de datos en tiempo real están asociados a modelos de decisión en línea que proporcionan recomendaciones tan pronto como se toma la decisión [40]. En este contexto, los recomendadores en línea desempeñan un papel esencial.

En [41], [42],los autores presentan diferentes escenarios aplicaciones donde el IoC representó experiencias exitosas, como el campo de la salud, donde es posible una mejora en la atención médica a través de la recolección de información en tiempo de sensores para toma de decisiones informadas. También destacan la importancia de una buena rehabilitación en personas con fracturas de cadera, utilizando el IoC en relación con la salud digital. De esta forma, se realiza el seguimiento de pacientes con dificultades en su capacidad motora. Permite evaluar y analizar el proceso de rehabilitación en cuanto llegan los datos de las fuentes de datos (es decir, los dispositivos montados en el paciente).

En [43], el IoC se aplica para predecir el tráfico de vehículos en ciudades inteligentes, a través de la recopilación de datos y la extracción de características que permiten optimizar el sistema de control de señales de tráfico. Es interesante porque permite coordinar los componentes móviles (por ejemplo los vehículos) junto con los dispositivos estáticos (ejemplo, los dispositivos de señalización) en busca de la optimización del flujo global.

Existe una serie numerosas de investigaciones relacionadas con conceptos como tiempo real o sistemas online donde se pueden observar diferentes aplicaciones. En [44], los autores proponen un modelo de predicción de datos para predecir y así poder detectar anomalías o errores humanos que puedan surgir en la planta nuclea, permitiendo el diagnóstico de accidentes en tiempo real.

La idea es que utilizando sistema de tecnología IoC, se pueda evaluar la cantidad de información que surge de la recolección de datos, para tomar decisiones en determinadas situaciones. Por eso es importante contar con una tecnología que ayude a dar soluciones en los diferentes campos o áreas donde se aplique. Sin embargo, al combinar conceptos como los sensores IoC y los sistemas en tiempo real, surgen aplicaciones que permitan actuar en un campo de acción más amplio, es decir, se pueden crear aplicaciones más complejas adaptándose a las necesidades de las personas. La idea de este tipo de aplicaciones es que se adapten a las necesidades de los usuarios y que puedan ser utilizadas en cualquier momento. Es posible encontrar aplicaciones relacionadas con la monitorización de la alineación de las ruedas [45], la monitorización de la salud [46], la detección de inundaciones [47], entre otras.

En esta sección, se aborda un análisis superficial centrado en sistemas de recomendación. Como se mencionó previamente, el método de investigación utilizado es conocido como SMS. El objetivo consiste en proporcionar una perspectiva actual y amplia sobre los sistemas de recomendación aplicados en entornos de tiempo real a través de dispositivos IoC. Las revisiones sistemáticas comienzan con la definición de un protocolo de revisión. En dicho protocolo se alinean el objetivo y las preguntas de investigación para guiar el estudio. A continuación, se describen las preguntas de investigación para esta sección y de detalla la aplicación del protocolo.

2.1.1 Definición de las preguntas de investigación

Un conjunto de preguntas guía la aplicación del protocolo en la revisión. Dichas preguntas se conocen como Preguntas de Investigación (PI - en inglés Research Questions). La definición de las preguntas de investigación es esencial porque determinan si el alcance de la revisión es amplio o estrecho. Al formular las preguntas de investigación, las mismas deben ser claras y precisas para que la búsqueda de los documentos sea precisa y fácil.

Las principales preguntas de investigación, se definen de la siguiente manera:

- **PI1:** ¿Qué tipos de recomendadores en tiempo real están asociados al IoC?
- **PI2:** ¿Cuáles son las áreas que más utilizan los sistemas de recomendación en tiempo real?
- **PI3:** ¿Cuáles son las técnicas utilizadas para analizar los datos llegados de los diferentes tipos de dispositivos del IoC?
- **PI4:** ¿Cuáles son los métodos (o marcos) utilizados para supervisar los datos en el entorno de la IoC?

- **PI5:** ¿Cuáles son los tipos de documentos que abordan los sistemas de recomendación en entornos en tiempo real?

Cada pregunta definida previamente junto con sus motivaciones, se muestran en la Tabla 2. Cada pregunta describe una perspectiva a través de la cual se analiza el tema. La columna de motivaciones describe la razón principal detrás del punto de vista y su alineación con el objetivo de la investigación definida.

Tabla 2 Preguntas y motivaciones que guían la investigación

Perspectiva	Motivaciones
PI1: ¿Qué tipos de recomendadores en tiempo real están asociados al IoC?	Analizar los tipos de recomendadores que puedan existir para evaluar, o que se aplican al tema tratado.
PI2: ¿Qué tipos de recomendadores en tiempo real están asociados al IoC?	Evaluar cuáles son las áreas donde los sistemas en tiempo real son más aplicables.
PI3: ¿Cuáles son las técnicas utilizadas para analizar los datos llegados de los diferentes tipos de dispositivos del IoC?	Analizar las diferentes técnicas que pueden existir para evaluar los datos de dispositivos heterogéneos de IoC.
PI4: ¿Cuáles son los métodos (o marcos) utilizados para supervisar los datos en el entorno de la IoC?	Determina si algunos métodos (o frameworks) diferentes permiten la monitorización de datos en el entorno de IoC. Así, evalúa cuál podría ser el más conveniente aplicable.
PI5: ¿Cuáles son los tipos de documentos que abordan los sistemas de recomendación en entornos en tiempo real?	Busca todas las publicaciones en las que esté presente el tema. Considera el tipo de publicación y su evolución en el tiempo.

En la Tabla 2 se describen las preguntas de investigación para guiar la revisión de la bibliografía junto con la descripción de las motivaciones planteadas de acuerdo al objetivo planteado.

Para realizar la revisión sistemática [48], es importante que se defina un protocolo para dar inicio al proceso de revisión. Los pasos para ejecutar el protocolo consisten en a) Definir los términos de la búsqueda; b) Identificar las bases de datos y los motores de búsqueda; c) Definir los criterios de inclusión y exclusión; d) Asegurar que los artículos seleccionados sean representativos y relevantes.

El SMS define una estrategia de búsqueda la cual se centra en detectar la mayor cantidad posible de documentos relevantes.

2.1.2 Especificación de la estrategia de búsqueda

Definidas las preguntas de investigación, es necesario describir la estrategia de búsqueda. La estrategia de búsqueda se especifica para que se reproducible. Esto

significa que cualquier investigador debería poder repetir el proceso en la bibliotecas digitales.

La Tabla 3 describe la consulta asociada a la estrategia de búsqueda. Se encuentra alineada con el objetivo de la revisión en conjunto con las motivaciones descritas anteriormente.

Tabla 3 Síntesis de la estrategia de búsqueda

Concepto principal	Conceptos alternativos
Real-time Recommender	(“Recommender” OR “recommendation” OR “recommending”) AND (“system”) AND (“Internet of Thing” OR “IoT” OR “IOT” OR “IIOT” OR “sensor”) AND (“real time” OR “real-time” OR “online”)

Definida la cadena de búsqueda (es decir, especificando los términos principales y alternativos, en conjunto con los operadores lógicos), es necesario especificar los parámetros de búsqueda ajustados a la biblioteca digital. En este caso, la cadena de búsqueda se ejecuta en la base de datos Scopus.

La Tabla 4 presenta la base de datos de destino junto con el tipo de artículos de interés. Se describen también el ámbito y el tiempo de búsqueda, definiendo un conjunto de parámetros necesarios para ejecutar la consulta.

Tabla 4 Una perspectiva basada en filtros para los resultados de la consulta

Base de Datos	Scopus
Tipo de artículos	Artículos, Revistas
Búsqueda aplicada a	Resumen, título o palabras claves
Idioma	Inglés
Período de tiempo	Sin limitaciones

Cada pregunta de investigación requiere una respuesta óptima desde el punto de vista metodológico. Por ello, es fundamental conocer qué bases de datos ofrecen la mejor adecuación a cada pregunta.

2.1.3 Proceso de selección de artículos

A continuación se detalla el proceso de selección de artículos. Se describe cómo se conservan los artículos que cumples los criterios de inclusión y se eliminan los que reúnen los criterios de exclusión.

La Tabla 5 muestra un resumen de la estrategia de la selección de artículos, describiendo los criterios de inclusión y exclusión.

Tabla 5 Criterios empleados para retener o excluir artículos

Retención	1. Elementos que satisfacen las restricciones de la consulta. 2. Los documentos deben estar escritos en lengua inglesa. 3. Sólo se consideran los documentos publicados en revistas. 4. No existe ninguna restricción en cuanto a la fecha de publicación.
Eliminación	1. Los artículos que no abordan directamente el IoC, el tiempo real y los recomendadores. 2. Los elementos que no satisfacen las restricciones de la consulta. 3. Sitios web.
Detalles adicionales de la eliminación	1. Los documentos que describen una síntesis, enfoque o resumen para el tema. 2. No se consideran las síntesis relacionadas con charlas o conferencias magistrales invitadas.

La selección de los artículos de la búsqueda, se siguen los siguientes pasos: a) Se realiza el análisis de documentos duplicados. Así, se retiene una copia única en aquellos casos en los que el documento aparece varias veces; b) Se filtran los resultados iniciales utilizando los criterios de retención y eliminación (Ver Tabla 5Tabla 5). De este modo se obtiene una lista depurada; c) La lista depurada se utiliza para leer en detalla cada documento contenido. Así, los documentos que abordan directamente el tema mencionado (es decir, permite eliminar los documentos que sólo mencionan el tema) componen la lista definitiva de documentos.

Cada artículo es evaluado de acuerdo con el objetivo de la revisión. Aquellos artículos cuyos contribuciones se ajustan a las preguntas de investigación y a los criterios de inclusión, se mantienen. Los demás artículos son eliminados de la lista. Es decir, ejecutada la cadena de búsqueda se obtiene la lista de resultados, de los cuales se revisa la idoneidad de los resúmenes (abstract), los títulos y las palabras claves. Con dicha lista depurada de acuerdo a los criterios definidos previamente, se procede a leer los documentos a texto completo obteniéndose la lista final de artículos para su comparación y análisis.

2.1.4 La perspectiva del procesamiento de datos

La lista final se organiza y procesa a partir de diferentes dimensiones alineadas con las perspectivas de investigación. Cada perspectiva de investigación y su motivación, pueden representar una dimensión. Así, cada dimensión puede tener asociada una o varias categorías.

En la Tabla 6 se puede observar la relación entre las preguntas de investigación, las dimensiones y las categorías. De este modo cada documento que forma parte de la lista final de artículos seleccionados, se leen y se clasifican considerando las dimensiones y categorías, como se pude observar a continuación.

Tabla 6 Síntesis de las dimensiones y categorías asociadas

Perspectiva	Dimensiones	Categorías
PI1	Tipos de recomendadores	Tiempo real, IoC
PI2	Área de aplicación	Tiempo real
PI3	Técnicas	IoC, tiempo real, recomendadores
PI4	Métodos	IoC, Tiempo real, sistemas
PI5	Publicaciones	Tiempo real, IoC, sistemas recomendadores

La Tabla 6 muestra las dimensiones y categorías de cada una de las preguntas de investigación definidas. Las dimensiones y categorías relacionadas con cada pregunta se determinan para poder desglosar los artículos encontrados en subgrupos. De este modo, la dimensión indica los conceptos generales utilizados en la revisión. La categoría establecidas para cada pregunta de investigación.

El establecimiento de las dimensiones y las categorías persigue evitar sesos en el proceso de selección, permitiendo ello que aumente la fiabilidad del proceso.

Es conveniente diseñar una hoja (en un archivo Excel) de recogida de datos donde se recojan todas las características del estudio como ser: autor, país, año, resumen, entre otros aspectos. Esto permite por un lado, armar una lista completa de artículos donde se incorporan todos los metadatos del documento, por ejemplo: título, resumen, autores, editorial, palabras claves, etc. Por otro lado permite describir los motivos por los que se mantienen o eliminan cada uno de los documentos de la lista.

Para la presente sección, se dispone de un archivo Excel, donde se listan los documentos encontrados.

2.1.5 Reducción de la lista de documentos

Al aplicar la metodología descripta, el proceso de síntesis es una parte fundamental. Reduce la lista mediante la aplicación de los criterios definidos. De este modo, controla la evaluación de cada artículo obtenido, realiza el análisis de duplicados y retiene sólo aquellos que satisfacen las condiciones y restricciones especificadas. Es decir, el resultado de la consulta original se reduce mediante la aplicación de diferentes criterios. Todos están alineados con el objetivo de la revisión. Todos los resúmenes de los artículos fueron leídos y se seleccionaron los que responden a las preguntas de investigación previamente definidas (véase la Tabla 2). A partir del resultado del paso anterior, se evalúan los resultados y los datos obtenidos para destacar los resultados de la revisión que son significativos y representativos.

2.1.6 Ejecución de la Revisión Sistemática de la Literatura

La ejecución de la revisión implica ejecutar una consulta definida y llevar a cabo cada uno de los pasos descritos anteriormente. La consulta especificada y descrita en la sección 2.1.2 (ver Tabla 3) permite encontrar los artículos a través de la base de datos Scopus. De este modo, la consulta se ejecuta en la base de datos Scopus para localizar artículos de revistas escritas en el idioma inglés dentro del tema específico. Por otro lado, se debe tener en cuenta que los documentos se retienen si y sólo si contienen los conceptos especificados (ver Tabla 2) en el resumen, el título o las palabras clave. En el paso siguiente se aplican los filtros y criterios de acuerdo con la Tabla 4 y Tabla 5.

La ejecución de la cadena de búsqueda proporciona una serie de documentos asociados al tema de investigación, en este caso se refiere a sistemas de recomendación. Los registros duplicaos se eliminan dejando sólo un artículo del número de elementos descartados. Se revisan los títulos, resúmenes y palabras claves de todos aquellos artículos que cumplan con el filtro.

La Figura 3, muestra una captura de pantalla correspondiente a la ejecución en el sitio web de la base de datos Scopus. La consulta se configura describiendo el objetivo de la revisión, según la sintaxis de Scopus.

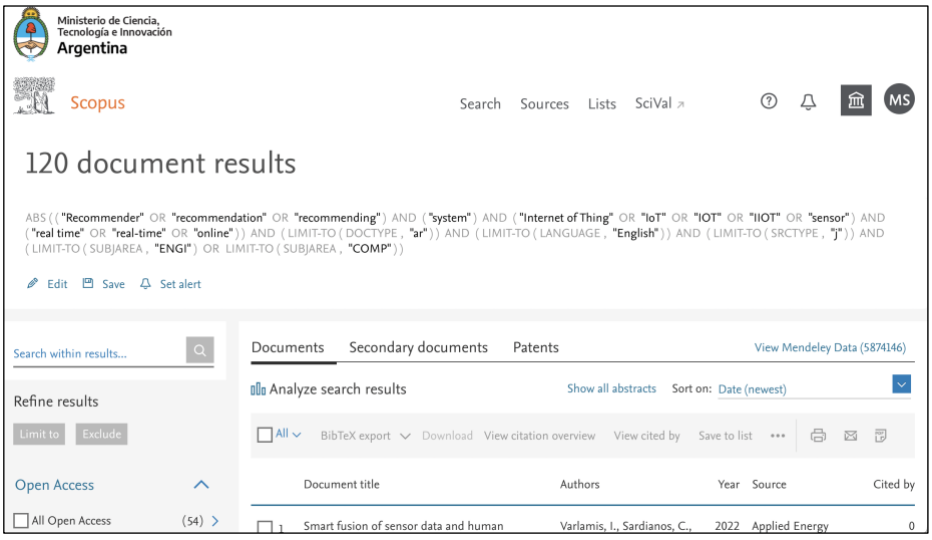

Figura 3 Resultados de la aplicación de la cadena de búsqueda. Captura de pantalla de la consulta ejecutada el 21-Julio-2021 a las 2:41 p.m. [49]

En la figura anterior se pueden ver los resultados obtenidos al ejecutar cadena de búsqueda. De la ejecución se obtienen 120 documentos. Se puede apreciar que los términos como ser (“Recomendador” O “recomendación” O “recomendando”) se encuentran incluidos en la cadena de búsqueda. En efecto, corresponden a aquellos conceptos alineados con el objetivo de la revisión que se espera se encuentren presentes en la lista de resultados. Sin embargo al refinar la cadena de búsqueda, y al aplicar en forma simultánea los filtros, la lista de resultados se reduce a 111 documentos. Se lee cada uno de los documentos. Se analiza cada situación. Es posible

eliminar un artículo mediante la lectura de solamente del resumen. Por otro lado, algunos artículos requieren ser leídos en forma completa.

Los artículos retenidos corresponden a aquellos documentos que satisfacen las condiciones y criterios especificados. De este modo se obtienen 12 documentos y se incorporan a la lista de documentos seleccionados.

2.1.7 Resumen de los resultados obtenidos

De la lista de selección, se conservan solamente 12 documentos. Luego de la lectura de los mismos, se puede indicar que los 12 documentos satisfacen las condiciones requeridas. Es decir, que los mismos son pertinentes para el análisis del tema y el objetivo de la revisión realizada.

En [50], los autores describen la aplicación de diferentes técnicas de aprendizaje automático en el ámbito sanitario. De este modo, utilizan conjuntos de datos públicos para construir un sistema centrado en la monitorización de la salud en tiempo real y a distancia. Sigue un enfoque tradicional basado en el IoC y la computación en la nube. Los datos se capturan del paciente a través de diferentes sensores y luego se transmiten a la nube para su análisis.

En [51], los autores introducen un análisis de patrones basado en dispositivos IoC, y una arquitectura Big Data. Como complemento, describen un esquema de recomendación que analiza el comportamiento de los datos para apoyar la toma de decisiones. De este modo, se capitalizan las experiencias y conocimientos previos, para organizar los cursos de recomendaciones.

En [52], los autores describen un análisis temporal de una entidad cuya monitorización se basa en dispositivos IoC. Se consideran lo sestados relacionados con la entidad bajo monitorización, las propiedades y los patrones de búsqueda para soportar una estrategia de recomendación colaborativa. Se implementa en la articulación entre la nube y Edge computing.

En [38], la propuesta describe un análisis de la estrategias de monitorización centradas en la monitorización de materia particulada basada en dispositivos de bajo costo. Se describen. Tecnologías actuales, las preocupaciones y las oportunidades , así como también el hardware y el software típicos utilizados con este fin. La propuesta describe el impacto de las partículas en la salud de las personas. Además, presenta los esfuerzos distribuidos para aumentar la resolución temporal y espacial de la monitorización de este pro9blema, basada en dispositivos IoC.

En [53], los autores proponen el desarrollo de una factura eléctrica en tiempo real para cuantificar la energía utilizada en instalaciones domésticas en México. Para ello, utilizan sensores de bajo costo y la placa electrónica Particle® Photon®. Su propuesta consiste en una herramienta de bajo costo que aprovecha la tecnología IoC para generar una factura de fácil lectura en tiempo real que permite a los clientes revisar y gestionar constantemente su consumo energético.

En [54], los autores describen una estrategia de recomendación basada en los casos de uso del sistema de promoción online de equipos IoC. Así, las recomendaciones se proporcionan de acuerdo con el comportamiento del usuario mientras utiliza el sistema online. De este modo, se persigue el análisis y la estimación de la demanda bajo una perspectiva de toma de decisiones basada en datos.

En [55], la propuesta analiza la estrategia de recomendación basada en eventos y contemplando un entorno de tiempo real. Así, se monitorizan un conjunto de elementos relacionados con los usuarios (por ejemplo, la ubicación). Esto permite proporcionar recomendaciones adecuadas y conscientes del contexto. Los autores describen una arquitectura de tres niveles, que combina e integra IoC, Edge Computing y Cloud Computing.

En [56], se describe una propuesta basada en servicios de IoC que se ejecutan en redes inalámbricas. En este contexto, un conjunto de dispositivos heterogéneos sirven datos de forma continua e ilimitada. Además, el receptor de los datos no tiene el control sobre el origen de los mismos (es decir, el propio dispositivo). En este entorno, el trabajo propone una estrategia de recomendación teniendo en cuenta los datos cuantificables y no cuantificables.

Por otro lado, en [57]; los autores describen una propuesta para abordar un aspecto normativo de la protección de datos. Aborda las preocupaciones relacionadas con la interoperabilidad sintáctica y semántica. Además, el esquema de recomendación proporciona rutas y espacios de estacionamiento para garantizar la privacidad del usuario cuando los datos deben ser transmitidos.

En [58], los autores proponen en su trabajo de investigación un sistema de detección de intrusos y de cualquier anomalía que se produzca durante el proceso de recogida de datos del banco de pruebas (Secure Water Treatment - SWat). Además, proponen autómatas temporizados basados en las variantes del proceso. Así, se introducen y describen patrones de ataque junto con el esquema de detección.

En [59], se describe una propuesta que aborda la monitorización del agua. Se centra en el nivel de calidad asociado al agua de un río. El grado de calidad se evalúa en función de requisitos específicos y normativos. Se describe una estrategia para implementar la monitorización activa en los recursos limitados, equilibrando las restricciones técnicas y los costos.

En [60], los autores proponen desarrollar un sistema habilitado por el Internet de las Cosas, que notifica el tiempo mínimo de desencofrado vertical en función de una resistencia a la comprensión del objetivo. Esto permite monitorizar la resistencia a la comprensión del hormigón de edad temprana en tiempo real, utilizando la tecnología IoC para determinar en tiempo mínimo requerido por la construcción.

Los artículos seleccionados describen el resultado de la aplicación del SMS. Permite responder en forma simultáneamente a cada una de las perspectivas detalladas en la Tabla 2. Todas las perspectivas responden al objetivo dela revisión según las dimensiones y categorías definidas. El volumen de documentos obtenidos indica un área

de investigación interesante. Se trata de un área en la que muchas aplicaciones son factibles debido al bajo costo asociado a las diferentes soluciones.

A continuación en la Tabla 7, se describe un análisis comparativo, contrastando cada trabajo presentado con las preguntas de investigación. Cada perspectiva está numerada del 1 al 5 en dicha tabla, siguiendo el mismo orden definido en la Tabla 2. Se incorpora además el número de citas recibidas junto con el año de publicación. El número de citas se obtiene de la base de datos Scopus. Se limita a ella hasta el día en que se ejecutó la consulta. La intersección entre un artículo determinado y una columna de preguntas de investigación específica, se indica con un espacio en blanco cuando no se ha encontrado ninguna relación. Sin embargo, cuando el documento proporciona una respuesta parcial o completa a la pregunta de investigación asociada, se indica con un símbolo de suma (es decir,+).

Los resultados contemplan artículos desde 2010 hasta ahora. Se analizan para el tema analizado. Sin embargo, hay un salto considerable entre 2010 y 2019. Justo en 2019 en adelante, se fusionan las contribuciones específicas en esta área. Podría asociarse al crecimiento y evolución de la oferta tecnológica en el tiempo.

Tabla 7 Relación entre las perspectivas y la lista de documentos depurados [49]

Artículos	Año	Citado	1	2	3	4	5
Experiences and recommendations in deploying a real-time, water quality monitoring system [59]	2010	58	+	+		+	+
A healthcare monitoring system using random forest and internet of things (IoT) [50]	2019	47		+	+	+	+
Big data and rule-based recommendation system in Internet of Things [51]	2019	10	+			+	+
Quantile context-aware social IoT service big data recommendation with D2D communication [56]	2020	6	+				+
Edge and cloud collaborative entity recommendation method towards the IoT search [52]	2020	3				+	+
Intrusion detection system using timed automata for cyber physical systems [58]	2019	2	+				+
Three-tier IoT-edge-cloud (3T-IEC) architectural paradigm for real-time event recommendation in event-based social networks [55]	2021	1		+	+		+
Iot-based electricity bill for domestic applications [53]	2020	1		+		+	+
IoTRec: The IoT Recommender for Smart Parking System [57]	2020	1	+	+			+
IoT-Based Approaches for Monitoring the Particulate Matter and Its Impact on Health [38]	2021	0	+	+			+
News Information Platform Optimization Based on the Internet of Things [54]	2021	0	+	+	+		+

An IoT device for striking of vertical concrete formwork [60]	2021	0	+	+			+

Podemos observar en la Tabla 7, que en forma simultánea ningún documento satisface las cinco perspectivas. Por otro lado, los documentos [50], [54], [59], llegan a responder a cuatro de las cinco perspectivas. Es más, el documento [50], representa el segundo documento más citado de la tabla. Mientras que con 58 citaciones se presenta el documento [59].

2.1.8 Resumen Final del SMS

Hasta aquí se describió la aplicación del SMS. Es importante mencionar que el punto de partida en esta metodología consiste en realizar una búsqueda exhaustiva, el análisis, la evaluación y la síntesis de documentos relacionados con un tema en particular.

Se definen y estructuran las preguntas de investigación para poder iniciar la revisión y guiar el estudio. En esta sección, se introdujeron términos principales y alternativos como “real-time”, “recommender”, “recommendation”, “recommended”, “system”, “internet of things”, “sensor” y “online” como parte del método de recopilación de información.

Se analizan los diferentes ámbitos de aplicación en los que los sensores basados en IoC y los sistemas de recomendación en tiempo real permiten dar respuesta a las necesidades básicas de información, permitiendo tomar decisiones en tiempo real en función de las diferentes situaciones planteadas.

En esta revisión, se obtuvieron 120 documentos. Se redujeron a través de los criterios de inclusión y exclusión según el protocolo de revisión. Se definió y aplicó un conjunto de filtros, se consideraron áreas asociadas a la informática, la ingeniería y la computación, entre otros aspectos. Del total de artículos analizados y considerando las preguntas de investigación planteadas, se concluyó que sólo 12 documentos cumplían simultáneamente con los requisitos establecidos en el método aplicado para realizar la revisión. Al momento de realizar una revisión sistemática, es importante tener en cuenta la calidad metodológica de los estudios incluidos en la revisión. De este modo, se evita el sesgo a la hora de elaborar la lista final de documentos seleccionados.

De acuerdo con los resultados obtenidos en esta revisión, el avance de la tecnología junto con las diferentes aplicaciones de los sensores IoC en áreas como la agricultura, la salud, la seguridad, entre otras, sigue siendo un tema a analizar. Los sistemas de recomendación en tiempo real y su aplicabilidad en la toma de decisiones dependen del contexto en el que están inmersos. Reunir las cinco perspectivas por las preguntas de investigación sigue siendo un reto, pero también una oportunidad.

Finalizado el SMS asociado a los sistemas de recomendación, se procederá a realizar un análisis de los diferentes tipos de marcos de medición y evaluación encontrados.

2.2 Marcos de Medición y Evaluación

Las organizaciones eficientes saben que todo aquello que no es posible medir o evaluar, tampoco se puede controlar. Existen determinadas situaciones en donde es necesario medir y evaluar los datos, lo que implica que si no se cuenta con la suficiente información sobre un determinado producto o proceso, no se podrá controla el mismo. En este sentido, es importante establecer qué se entiende por medición y qué se entiende por evaluación. Para ello, podemos definir a la medición como la asignación de un valor o el grado de cantidad de algo, cuya característica es que debe ser exacta y fiable, de manera que pueda proporcionar resultados de calidad. Las mediciones bien diseñadas, nos permiten comparar los resultados obtenidos con los resultados esperados. Por otro lado, se entiende por evaluación, al proceso sistemático de identificar, recolectar o bien tratar datos para luego tomar decisiones de acuerdo a la valoración realizada sobre los mismos. Se considera a la evaluación como el punto de partida para llevar a cabo la evaluación de los resultados y de este modo poder comparar los mismos. Al momento de utilizar ambos términos, debemos tener en cuenta cuáles son las diferencias más significativas entre ellos y considerar que ambos términos no son sinónimos entre sí. De este modo, y a fin de evitar confusiones o errores a la hora de aplicar los mismos es necesario diferenciar las características de cada uno de ellos.

A continuación en la Tabla 8, se indican las principales diferencias entre los términos.

Tabla 8 Principales diferencias entre Medición y Evaluación

Medición	Evaluación
• Es la base de la evaluación • Tiende a cuantificar • Se centra en objetivos pre establecidos	• Se apoya en la medición • Tiende a cualificar • Se centra en la persona y a partir de allí da una valoración

Si bien existen diferencias significativas entre ambos términos, como se pudo observar en la tabla anterior, debemos tener en cuenta que ambas persiguen el mismo propósito que consiste en la toma de decisiones, tal como se muestra en la Figura 4:

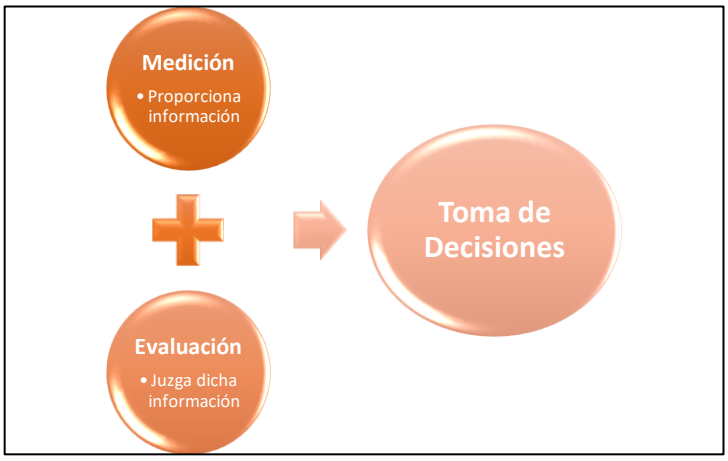

Figura 4 Objetivo Principal de Medición y Evaluación

Teniendo en cuenta las diferencias y características de los conceptos antes mencionado, debemos considerar que cuando es necesario tener que definir un proyecto de medición y evaluación, lo que se suele utilizar es un marco formal de medición y evaluación. Ello permite tener en cuenta la entidad bajo monitoreo que se encuentra analizando.

Dentro de los marcos formales de medición, podemos mencionar al "framework" denominado CONSERVE [61], cuyo objetivo principal es permitir la toma de decisiones multidimensional de un conjunto de técnicas de monitoreo en entornos de virtualización para supervisar la seguridad de los contenedores. La idea principal es utilizar dicho "framework" para mejorar la eficiencia y fiabilidad de los entornos de ejecución. Por otro lado, para poder estimar la calidad de un producto en forma directa, a partir de los diferentes datos provenientes del proceso de producción, es que se aplica la metrología virtual. Sin embargo, es necesario contar con un "framework" que permita estructurar todas las actividades de investigación relacionadas con la metrología virtual [62]. De este modo, es posible proporcionar un marco y terminología que fomente la investigación y el control en tiempo real de todos los componentes y aplicaciones de la metrología virtual. Por otro lado, si se desea medir y evaluar el rendimiento de un sistema de transporte de mercancía [63], es necesario contar con un "framework" que permita realizar dicha medición a través de la incorporación de elementos numéricos que posibiliten la toma de decisiones basado en conjuntos aproximados. El objetivo de dicho marco es evaluar su aplicabilidad en el análisis del rendimiento de empresas que transportan mercancía permitiendo su comparación con otros métodos y de este modo validar la eficacia del enfoque propuesto.

En este sentido, cabe destacar que los diferentes tipos de "framework" nos permiten mejorar y optimizar la eficacia de las distintas aplicaciones donde podemos utilizarlos, con el objetivo de brindar una estructura que sirva de base para los proyectos a desarrollar.

2.2.1 Marco Conceptual C-INCAMI

Cuando es necesario medir y evaluar un ente, el mismo debe ser definido y delimitado, de este modo se podrá conocer el objetivo y el alcance del mismo[64].

Establecer una estrategia de medición y evaluación sobre un proceso, significa que se debe poder estudiar su comportamiento, poder identificar y detectar anomalías, como así también poder evaluar su resultado. Sin embargo, la distancia entre la formalización de un proceso hasta la visualización de los indicadores que permitan tomar decisiones sobre dicho proceso, implica la utilización de un marco formal de medición y evaluación (con base ontológica). Ello permitirá, garantizar la consistencia, la repetividad y la comparabilidad de las medidas en el tiempo.

El marco C-INCAMI (en inglés Context-Information Need, Concept Model, Attribute, Metric and Indicator) [2], [7], [65] consiste en un marco conceptual que define módulos y conceptos que intervienen en el área de la medición y evaluación. El mismo, se basa en un enfoque donde las especificaciones de los requerimientos, la medición y la evaluación de entidades, como así también la interpretación de los resultados obtenidos, se encuentran orientados a satisfacer una necesidad de información en particular.

C-INCAMI se encuentra integrado por los siguientes componentes:

- Definición del proyecto de medición y evaluación
- Definición y especificación de requerimientos no funcionales
- Especificación del contexto del proyecto
- Diseño y ejecución de la medición
- Diseño y ejecución de la evaluación
- Especificación del análisis y recomendación

Los componentes se encuentran soportados por los términos ontológicos definidos en [2], [7], [65] y en la Figura 5, se puede observar los principales conceptos y relaciones para la especificación de requerimientos no funcionales, la especificación del contexto, diseño e implementación de la medición, y diseño e implementación de la evaluación.

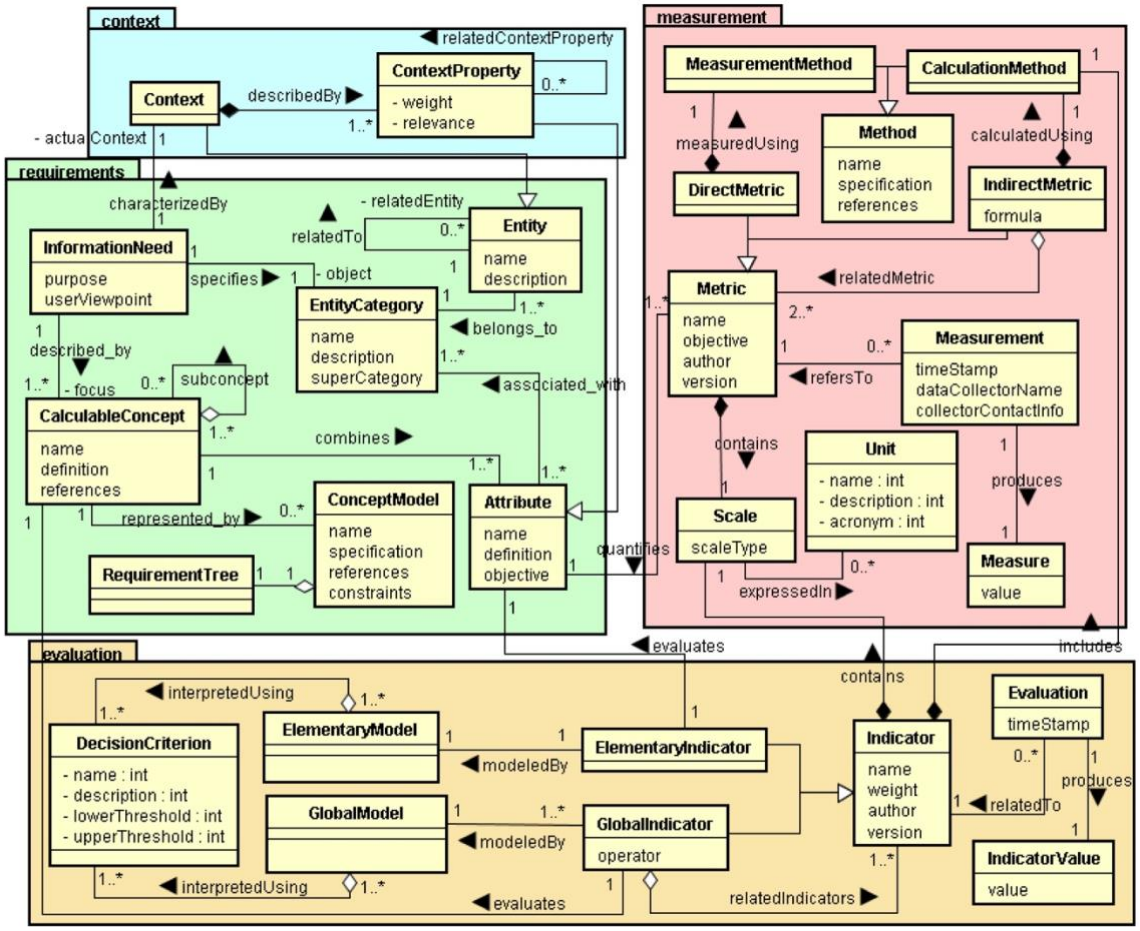

Figura 5 Principales conceptos y relaciones de C-INCAMI [66]

El componente denominado *Definición del proyecto de medición y evaluación* (no se muestra en Figura 5), permite definir el proyecto sobre el cual luego se especificarán los requerimientos no funcionales asociados a las actividades de medición y evaluación. Este componentes, define y relaciona un conjunto de conceptos del proyecto que son necesarios para tratar las actividades, los roles y los artefactos de M&E. En lo que respecta a los datos asociados a la definición del proyecto, podemos encontrar su nombre o denominación, período de vigencia, director, entre otros datos.

En cuanto al componente *Definición y especificación de requerimientos no funcionales* (paquete de requisitos en la Figura 5), el mismo permite especificar la necesidad de la información de un proyecto de medición y evaluación. La necesidad de información describe el objetivo o propósito (entender, predecir, controlar, etc) y el punto de vista del usuario (paciente, usuario final, etc). También identifica un concepto calculable, por ejemplo, calidad del software y especifica la categoría de las entidades a evaluar; por ejemplo, un producto. Un concepto calculable, (característica o dimensión), puede ser definido como una relación abstracta entre atributos de una entidad y una necesidad de información; el cual puede ser representado por un modelo.

El componente *Especificación del contexto del proyecto* (paquete de contexto en la Figura 5), permite describir de forma estructurada el contexto relevante para una necesidad de información por medio de propiedades. Un concepto es *Contexto,* que

representa el estado relevante de la situación de la entidad a evaluar con respecto a la necesidad de información. Se considera el contexto como un tipo especial entidad en el que están involucradas entidades relevantes relacionadas. Para describir el contexto se utilizan atributos de las entidades relevantes.

El componente denominado *Diseño y ejecución de la medición* (paquete de medición en la Figura 5), permite especificar las métricas utilizadas en la medición como así también permite registrar los valores medidos de los atributos de cada entidad. Este componente incluye los conceptos y relaciones destinados a especificar el diseño y la implementación de la medición. Cabe mencionar que un atributo puede ser cuantificado por muchas métricas, pero un determinado proyecto de M&E solamente se utiliza una métrica para su cuantificación. La métrica define el método de medición o cálculo para obtener el valor del atributo y la escala de los valores. Un método de medición se aplica a una métrica directa, mientras que un método de cálculo se aplica a una métrica directa. Seleccionadas las métricas, se utiliza su definición para efectuar la medición y de este modo producir una medida para cada atributo. Sin embargo el valor obtenido por una métrica no representa el nivel de satisfacción de un requerimiento elemental (atributo) sino que es necesario realizar una nueva correspondencia utilizando indicadores. Dos tipos de métricas se especifican. Por un lado la métrica denominada *Métrica Directa,* que es aquella cuyo valores son obtenidos directamente de la medición del atributo de la entidad correspondiente. Por otro lado, se encuentra definida la *Métrica Indirecta* la cual se calcula a partir de los valores de otras métricas directas siguiendo una especificación de función y un método de cálculo determinado.

El componente definido como *Diseño y ejecución de la evaluación* (paquete de evaluación en la Figura 5), permite especificar los indicadores que permiten interpretar los atributos y conceptos calculables. Este componente incluye los conceptos y relaciones destinados a especificar el diseño y la implementación de la evaluación. Podemos mencionar que existen dos tipos de indicadores: *elementales y globales*. Un indicador se define como *elemental* dado que es aquel que no depende de otros indicadores para evaluar o estimar un concepto de bajo nivel de abstracción como son los atributos. Cada indicador elemental, tiene un modelo elemental que proporciona una función de mapeo de las medidas de las métricas. Por otro lado, un indicador *parcial o global*, se deriva de otros indicadores para poder evaluar o estimar un concepto de alto nivel de abstracción (conceptos y calculables y subconceptos). El valor del *indicador global* representa el grado de satisfacción de los requerimientos para dicha necesidad de información.

Finalmente el componente definido como *Especificación del análisis y recomendación* (que no aparece en la Figura 5), debe soportar un proceso en el que se diseña el análisis, se implementa el mismo y en base a sus resultados se elaboran reportes que posibiliten luego la elaboración de las recomendaciones [67]. Incluye conceptos y relaciones que tratan el diseño y la implementación del análisis, así como también la conclusión y la recomendación. El análisis y la recomendación utilizan información que proviene de

cada proyecto de M&E, ello incluye requisitos, contexto, datos de medición y evaluación, y metadatos.

En este sentido, la idea del marco es fomentar la repetibilidad, comparabilidad, coherencia y extensibilidad del proceso de medición. Por otra parte, la especificación de requisitos, la M&E y el análisis de resultados deben satisfacer una necesidad de información específica, en un contexto determinado. Los conceptos y relaciones de C-INCAMI (por ejemplo, Medida, Entidad, Contexto, etc.) están pensados para ser utilizados a lo largo de todas las actividades de M&E. De este modo, se pretende obtener una comprensión común de los datos y metadatos compartidos entre proyectos para fomentar un análisis más coherente.

2.3 Medidas de Similitud

Los criterios de decisión utilizan el conocimiento y las experiencias previas establecidas por los expertos en cada proyecto de medición para interpretar un valor numérico dado (es decir, una medida) a través de los indicadores para obtener un clasificador adecuado que proporcione recomendaciones. Por un lado la métrica se encarga de obtener el valor numérico. Por otro lado, el indicador está orientado a interpretar el valor, utilizando criterios de decisión. Los criterios de decisión están relacionados con los sistemas de recomendación que proporcionan un conjunto de posibles cursos de acción a través de clasificadores [68].

Cuando un proyecto de medición es nuevo o hay poca (o ninguna) información de cierta situación, no existe conocimiento previo para obtener un clasificador adecuado en el proyecto actual, por lo que podría derivar en una recomendación vacía.

Desde la perspectiva de la teoría de la información, dos conceptos son similares cuando comparten una serie de propiedades comunes. Así, la similitud podría representarse como una magnitud entre 0 y 1. Cuando un valor tiende a estar cerca de 1, ambos conceptos tienden a ser similares. Sin embargo, cuando el valor se acerca a 0, tienden a ser diferentes. Pero la similitud semántica se refiere a considerar el significado del contenido para aproximar dicha similitud y no la mera intersección de propiedades. Existen diferentes métodos para aproximar y calcular similitudes [69].

Dado que las definiciones de los proyectos de medición se establecen sobre una base conceptual común apoyada en una ontología C-INCAMI [70], el significado subyacente entre los conceptos estaba previamente definido.

Así, conociendo todas las definiciones de proyectos de medición, el problema reside en cómo encontrar proyectos de medición similares basados en el “framework” de medición para describir una ruta de búsqueda de nuevas recomendaciones guiada por la similitud entre proyectos. En otras palabras, dada una definición de proyecto, la idea es cómo obtener una lista de proyectos ordenadas pro similitud descendente para guiar el plan de recomendaciones. Anteriormente, se propuso un índice de similitud conductual y estructural [16], capitalizando la ontología C-INCAMI. Así, cuando se dispone de información, el índice era útil para llegar a definiciones de proyectos

similares, permitiendo explorar en clasificadores similares para obtener cursos de acción, llegando a una recomendación lo más cercana posible a la situación actual.

En la realidad virtual, la medición de la entidad objeto de seguimiento y el contexto son esenciales para comprender la relación e influencia con el mundo real [71]. Así una entidad podría experimentar diferentes estados, mientras que el contexto podría transitar por diferentes escenarios. Además, una medida determinada podría interpretarse de forma diferente cuando una entidad se encuentra en un estado determinado realizando alguna actividad en un escenario determinado. Por ello, en [16]se propuso una extensión de la ontología de medidas, incorporando escenarios conjuntamente con estados para modelar el dinamismo de la categoría de contexto y entidad respectivamente. Esto permite analizar entidades y contextos de diferentes perspectivas para conocer la similitud. Sin embargo, el índice de similitud original no contempla los escenarios y los estados de las entidades para guiar la búsqueda de recomendaciones similares ante nuevos proyectos.

Cuando una situación es nueva, puede que no existan cursos de acción (o clasificador) relacionados con la interpretación de los indicadores. Podemos considerar que cuando un proyecto dado no tiene cursos de acción (o clasificadores relacionados) para un criterio de decisión, un orden inteligente basado en la similitud podría guiar la búsqueda de clasificadores pertinentes entre otros proyectos de medición heterogéneos para mitigar el riesgo de arranque en frío.

El concepto de similitud está asociado a la medición de la proximidad de dos conceptos (es decir, palabras, entidades, etc.) a través de diferentes perspectivas [69], [72]. En esta sección, se describe por un lado la similitud semántica basada en la ontología y por otro lado, se sintetizan las estrategias de análisis basadas en la similitud articuladas al filtrado colaborativo.

2.3.1 Similitud Semántica basada en la ontología

En [73], los autores introducen un método para la similitud semántica basada en ontologías para estimar la fuerza de la relación entre dos entidades. Como punto en común, la similitud semántica se estima en base a ontologías, pero la diferencia radica en que los autores comparan la similitud semántica entre diferentes ontologías. En el presente trabajo de tesis, se plantea que los proyectos de medición son definidos en base a una misma ontología de medición, lo que permite caracterizar una entidad en términos de sus atributos, contextos y propiedades de contexto. De este modo, la similitud compara dos proyectos de medición diferentes basados en la misma ontología. Esto permite contemplar la perspectiva estructural y semántica en forma conjunta con el comportamiento de los datos a través de sus respectivas distribuciones de datos para las métricas asociadas.

En [74] se propone un método basado en características para estimar la similitud semántica entre ontologías. Cuando algún concepto no es lo suficientemente preciso o falta en una ontología dada, la idea es buscar ontologías similares para complementarla.

Como punto compartido, este trabajo produce un alista de basada en la similitud con el objetivo de buscar clasificadores similares en proyectos activos. Ello permite contemplar los clasificadores que faltan en un proyecto actual con clasificadores prevenientes de proyectos similares. La diferencia con nuestra propuesta, se manifiesta en que los proyectos de medición son descriptos utilizando la misma ontología de medición, sin embargo la instanciación de cada proyecto de medición tiene diferentes objetivos y entidades. De este modo, el índice compuesto se encarga de proporcionar la lista de proyectos activos ordenados por similitud a fin de guiar la búsqueda de recomendaciones.

En [75] Smaili et al introducen un método denominado OPA2Vec (en inglés Ontologies Plus Annotations to Vectors) para expresar una entidad biológica como un vector. El método combina axiomas de ontología y axiomas de anotación de los metadatos de la ontología para utilizar una estrategia Model2Vec. OPA2Vec podría utilizarse para producir representaciones vectoriales de una entidad biomédica. La perspectiva en común, las anotaciones basadas en al ontología que guían el procesamiento de los datos para alcanzar un valor de similitud. La similitud se emplea para ordenar la recomendación en el resultado. Como punto de diferencia, este trabajo se encuentra orientado a proyectos de medición donde la fuente de datos son sensores heterogéneos. Las entidades y sus contextos se definen en el proyecto de medición a partir de una ontología de medición. Los criterios de decisión de los indicadores que interpretan los datos procedentes de los sensores podrían adolecer de la ausencia de un clasificador que proporcione recomendaciones (es decir, un reto de arranque en frío). Para mitigar esta situación, un índice compuesto calcula la similitud entre proyectos para proporcionar una lista de lectura de proyectos basados en la similitud que orienta la búsqueda de clasificadores similares para ser utilizados temporalmente.

2.3.2 Estrategias de análisis basada en la similitud

En [76] Chung et al. proponen una estrategia basada en el entorno para la evaluación del riesgo para la salud. Los datos médicos de los pacientes se recopilan, se procesan y se unen a los datos ambientales. Mediante un motor de inferencia ontológica los datos y sus contextos se presentan como metadatos ontológicos. La similitud entre los pacientes se calcula mediante la fórmula de la distancia de Minkowski. Así, el modelo propone una evaluación del riesgo contrastando el índice de similitud de una persona con un patrón de una enfermedad crónica. Como perspectiva compartida, el índice considera factores externos e internos para obtener un valor. Como diferencia, la ontología subyacente se especializa aquí en el proceso de medición, describiendo la posibilidad de interpretar escenarios y estados de la entidad conjuntamente con la interacción entre ellos. Esta propuesta busca similitudes entre diferentes proyectos de medición para proporcionar una lista de lectura de proyectos basada en la similitud para guiar la búsqueda de recomendaciones. Permite ordenar el plan de consulta en el espacio de búsqueda de recomendaciones teniendo en cuenta todos los proyectos de

monitorización activos. Incluso, este cálculo de similitudes podría ser realizado por dispositivos IoC de bajo coste antes de solicitar recomendaciones, sugiriendo una ruta de consulta.

En [77], los autores describen una estrategia para proporcionar una lista ordenada de recomendaciones basada en datos históricos en sistemas recomendadores de filtrado colaborativo. La propuesta se centra en evaluar la posible lista de recomendaciones que mejor se ajusta al usuario activo desde una perspectiva semántica. Como punto en común, una decisión (en este caso una lista de recomendaciones) trata de ajustarse al usuario activo dada la información semántica sobre el usuario y su comportamiento. En nuestra propuesta, al descripción del usuario se obtiene mediante la especificación de la categoría de la entidad en el proyecto de medición basado en la ontología de medición. El comportamiento se aborda a través de medidas relacionadas con las métricas que cuantifican los atributos de la entidad o las propiedades del contexto. Como diferencia, esta propuesta proporciona una lista de lectura basada en la similitud para los proyectos de medición activos. Así, cuando no se asocia ningún clasificador a los criterios de decisión, se obtiene un clasificador similar guiado por una lista basada en la similitud.

En [78], se propone una propuesta de filtrado colaborativo basada en correlación. El índice de correlación entre ítems se analiza bajo el paragüas de la similitud semántica y contemplando una puntuación global para cada atributo. Así, el nivel de asociación lineal entre la similitud semántica y la puntuación global permite calcular la correlación entre ítems empleada por el algoritmo para proporcionar recomendaciones. Como perspectiva compartida, el nivel de asociación entre conceptos basado en etiquetas semánticas guía la estrategia de búsqueda. Como diferencia, nuestra propuesta se centra en analizar las similitudes entre proyectos de medición heterogéneos para llegar a una lista basada en la similitud. Esta lista es útil para guiar la búsqueda de clasificadores similares en proyectos activos y heterogéneos cuando no hay ningún clasificador disponible en el proyecto actual.

A continuación se aborda la aplicación del método de investigación conocido como SMS [19], donde el objetivo es brindar respuesta a la siguiente cuestión: ¿Es posible detectar entidades semánticamente similares en un proyecto de medición y evaluación (M&E) basado en el marco de medición C-INCAMI?

Tal y como se ha indicado previamente, esta sección será desarrollada siguiendo las directrices del SMS especificadas en [20]–[23] se aplica al procesamiento de flujo de datos y a la ciencia de datos.

El tema de las entidades semánticamente similares a la toma de decisión en tiempo real, no es un aspecto común o frecuente. Sin embargo, la aplicación del mapeo sistemático en el aprendizaje automático no es nuevo, por ejemplo, el mapeo se ha aplicado en la minería de procesos [79], en datos de ingeniería de software [80], el aprendizaje automático aplicado las pruebas de software [81], entre otros.

En esta sección se procederá a detallar el protocolo de revisión considerando los siguientes pasos: a) Identificar la necesidad de revisión, b) Especificar las preguntas de investigación, c) Determinar la estrategia de búsqueda, d) Especificar el proceso de extracción de datos, e) Especificar el proceso de síntesis y f) Revisar el desarrollo del protocolo. El protocolo guía la revisión para determinar los criterios de selección de los artículos encontrados. Las preguntas de investigación se describen a continuación.

2.3.3 Especificación de las preguntas de investigación

El objetivo en esta sección está relacionado con la realización de un mapeo sistemático de la literatura, con el fin de encontrar técnicas y métodos que permiten detectar entidades semánticamente similares en procesos de medición.

De acuerdo con la metodología utilizada, se definen las preguntas de investigación (PI – en inglés Research Question) que guían la investigación:

- **PI1:** ¿Qué métodos existen para detectar entidades semánticamente similares en un proyecto de M&E?
- **PI2:** ¿Qué tipos de entidades se comparan para ver su similitud?
- **PI3:** ¿Qué tipos de similitud se pueden encontrar entre entidades?
- **PI4:** ¿Qué tipo de publicaciones han tratado el tema de similitud semántica a lo largo del tiempo?

La Tabla 9 presenta cada pregunta de investigación, describiendo su definición, junto con las motivaciones asociada a cada una de ellas.

Tabla 9 Motivaciones asociadas a las preguntas de investigación

Preguntas de Investigación	Motivaciones
PI1: ¿ Qué métodos existen para detectar entidades semánticamente similares en un proyecto de M&E	Determinar si existen diferentes métodos que permitan la detección de entidades similares y así evaluar cuál podría ser el más conveniente aplicable al proyecto de M&E utilizado.
PI2: ¿Qué tipos de entidades se comparan para ver su similitud?	Evaluar cuáles son los tipos de entidades, y de éstas determinar si existe una similitud entre ellos.
PI3: ¿Qué tipos de similitud se pueden encontrar entre entidades?	Analizar los tipos de similitud que pueden existir para evaluar cuál o cuáles son aplicables al tema tratado.
PI4: ¿Qué tipo de publicaciones han tratado el tema de similitud semántica a lo largo del tiempo?	Consultar en todas las publicaciones en las que el tema ha sido presentado, teniendo en cuenta el tipo de publicación y su evolución en el tiempo.

Una vez que fueron descritas las preguntas de investigación y sus motivaciones, realizar el mapeo sistemático [48], implica definir el protocolo para dar inicio al proceso de revisión.

Para ello a continuación, se define la estrategia de búsqueda asociada al SMS, la cual se centra en detectar documentos que sean relevantes.

2.3.4 Definición de la estrategia de búsqueda

Al establecer la estrategia de búsqueda, que permite llevar a cabo la investigación y guiar el proceso, es importante considerar que la misma debe ser reproducible. Esto significa que se debería poder repetir el proceso en bibliotecas digitales que el investigador considere.

La Tabla 10 muestra la consulta asociada a la estrategia de búsqueda. La misma debe estar alineada con el objetivo de la investigación en forma conjunta con las descripciones descritas previamente.

Tabla 10 Descripción de la cadena de búsqueda

Término principal	Términos alternativos
Similarity Semantic	((“similarity semantic”) OR (“coefficient similarity”) OR (“measure similarity”) OR (“similarity between text”) OR (“semantic measurement text”))

Definida la cadena de búsqueda, es necesario identificar y definir la estrategia asociada a la misma. En la Tabla 11 se resume la estrategia de búsqueda, estableciendo los parámetros a considerar.

Tabla 11 Resumen de la estrategia de búsqueda

Base de Datos	Scopus
Tipo de artículos	Artículos, Artículos de Revistas, Documentos de conferencias, Capítulos de libros
Búsqueda aplicada a	Resumen, título o palabras claves
Idioma	Inglés
Período de tiempo	Todos hasta el presente

La base de datos seleccionada para llevar a cabo la estrategia de búsqueda fue Scopus. En la misma se ejecutó la cadena de búsqueda definida previamente, considerando como resultados destacados los artículos y revistas.

2.3.5 Proceso de selección de artículos

Luego de establecida la estrategia de búsqueda, se procede a detallar el proceso de selección de artículos, donde se determinan aquellos artículos que se deben considerar.

La Tabla 12 muestra un resumen de la estrategia de selección de artículos indicando los criterios de inclusión y exclusión que deben estar alineados con la metodología seleccionada para llevar a cabo el mapeo.

Tabla 12 Resumen de la estrategia de selección de artículos

Criterios de inclusión	1. Términos que se ajustan a la cadena de búsqueda 2. Publicaciones escritas en inglés 3. Artículos, revistas, documentos de conferencias, capítulos de libros 4. Fecha de publicación: todas hasta hoy
Criterios de exclusión para el título y el Abstract	1. Documentos que no centran en la informática 2. Artículos que no cumplan con la cadena de búsqueda 3. Blogs personales
Criterios de exclusión para el texto completo	1. Publicaciones que presentan un resumen de alguna charla 2. Uso de la similitud semántica en otras áreas diferentes a la informática

Al momento de seleccionar los artículos, los pasos realizados fueron los siguientes: a) Se eliminaron los artículos duplicados; b) Los artículos se filtraron utilizando los resultados del paso anterior aplicando los criterios de inclusión y exclusión establecidos en la Tabla 12; c) A partir de los resultados filtrados, se procedió a la lectura de cada artículo para seleccionar los que integrarían la lista definitiva de la investigación.

2.3.6 Proceso de extracción de datos

A partir de los artículos obtenidos en el proceso de selección (es decir la lista definitiva de artículos para la investigación), se prepara el formulario de extracción de datos, indicando el esquema de clasificación, detallando las dimensiones correspondientes, y las categorías asociadas. (Ver Tabla 13)

Tabla 13 Detalle de las dimensiones y categorías

Preguntas de investigación	Dimensiones	Categorías
PI1	Métodos	Entidades, similar, semánticamente
PI2	Tipo de entidades	Entidades, similares
PI3	Tipo de similaridad	Entidades, similitud
PI4	Publicación	Similitud, semántica

En la tabla anterior se puede visualizar las dimensiones y categorías asociadas a cada una de las preguntas de investigación, las cuales se definen para poder agrupar los artículos encontrados en subgrupos, y de esta manera permitir su análisis.

Como material de apoyo, se preparó un archivo de Excel incorporando todos los artículos encontrados e indicando las razones para incluir o excluir cada uno en base a la lectura del resumen/artículo completo.

2.3.7 Proceso de síntesis

Este proceso forma parte integral de la metodología utilizada. Se encarga de evaluar cada artículo encontrado, eliminando aquellos duplicados y seleccionando sólo los que cumplan con la cadena de búsqueda, los criterios de inclusión y exclusión, las palabras claves indicadas como así también los tipos de documentos filtrados.

Seguidamente, los resúmenes de cada uno de los artículos descargados fueron leídos y seleccionados aquellos que respondían en forma parcial o total a las preguntas de investigación definidas en la Tabla 9.

2.3.8 Ejecución del SMS

La ejecución del protocolo se refiere a realizar la búsqueda de acuerdo a la cadena de búsqueda establecida en la Tabla 10, así como la estrategia de búsqueda especificada en la Tabla 11, ello permite encontrar los artículos a través de la base de datos Scopus.

La Figura 6, muestra una captura de pantalla relacionada con la ejecución de la cadena de búsqueda definida previamente, en el sitio web de la base de datos utilizada, en este caso corresponde a la base de datos Scopus.

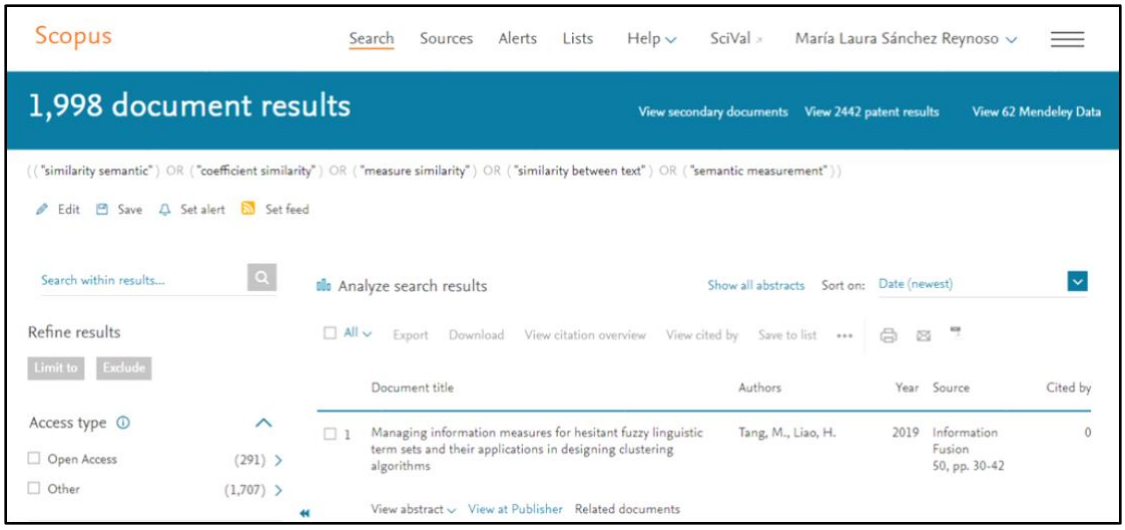

Figura 6 Captura de pantalla relacionada con el número de documentos encontrados al ejecutar la cadena de búsqueda el 26 Febrero 2019 a las 14:42 hs [82]

De la aplicación de la cadena de búsqueda, se obtuvieron 1998 registros. Sin embargo, al incorporar (“entidad” OR “entidades”) al principio de la cadena de búsqueda, permite refinar la consulta estableciendo que los términos indicados previamente, deben estar presentes en los documentos listados como resultado.

La cadena de búsqueda refinada, da como resultado 106 documentos que se filtran como i) Área temática: se limita a las áreas de “informática” e “ingeniería”; ii) Tipo de

documento: sólo se consideran las ponencias de conferencias, artículos y los capítulos de libros; iii) Palabras claves: se elige “semántica”, “similitud semántica” y “similitud”; iv) Idioma: inglés. Una vez aplicados los filtros, en forma simultánea sólo se obtuvieron 38 documentos. De esos 38 documentos, la selección de los artículos se realizó mediante la lectura del resumen (abstract) considerando los principales y alternativos, conjuntamente con las preguntas de investigación. De este modo se obtuvieron 4 documentos entre 2010 y 2018, correspondiendo todos ellos a artículos de revistas.

2.3.9 Resumen de los resultados obtenidos

De la lista final relacionada con los documentos seleccionados, sólo cuatro satisfacen en forma simultánea los requisitos relacionados con los filtros, la pertinencia y las preguntas de investigación.

En el primer trabajo [83], los autores describen un proceso organizado en tres etapas para medir y comparar ontologías. En este sentido, se introduce el concepto de medida semántica estable. Incluso cuando la similitud semántica se aplica a las ontologías, la idea subyacente podría estar relacionada con la entidad bajo monitoreo como forma de determinar la similitud semántica entre ellas.

En el segundo trabajo [84], los autores proponen un enfoque no supervisado para medir similitud semántica en textos cortos. Es muy interesante porque la descripción es una alternativa para incorporar en el momento en que se defina una entidad.

En el tercer trabajo [85], el autor propone *ColorSim,* una forma de cuantificar la similitud semántica teniendo en cuenta la semántica embebida en las anotaciones de OWL 2 (en inglés Web Ontology Language).

En el cuarto trabajo [86], los autores introducen una nueva técnica para calcular la similitud semántica entre dos entidades. Es muy interesante porque la técnica se basa en el conocimiento estructurado procedente de una ontología o taxonomía. La técnica propone el uso de los conceptos multi-árbol para determinar su similitud basándose en la ontología o taxonomía subyacente.

Los artículos seleccionados permiten dar respuesta simultáneamente cada una de las preguntas de investigación alineadas con el objetivo del presente SMS. El número de artículos obtenidos no es elevado pero los resultados son pertinentes para la línea de investigación y útiles desde el punto de vista conceptual.

La idea subyacente relacionada con las entidades semánticamente similares es reutilizar el conocimiento y las experiencias previas cuando alguna entidad similar no tiene evidencia o es muy limitada.

En este mapeo se encontraron como se mencionó, un total de 1998 documentos, reducidos a 106 al aplicar los criterios de exclusión y análisis. Se aplicó un conjunto de filtros relacionados a la informática e ingeniería, entre otros aspectos. Finalmente sólo cuatro trabajos satisfacían en forma simultánea los requisitos asociados al objetivo de la metodología aplicada. La similitud semántica aplicada a la entidades objeto de seguimiento en los proyectos de medición y evaluación, es un desafío. La toma de

decisiones en tiempo real depende de las medidas obtenidas, la entidad que se monitorea y del contexto en el cual está inmersa. De este modo, mediante la reutilización del conocimiento y las experiencias previas, un sistema de recomendación basado en memoria organizativa podría ofrecer un curso de acción análogo para la toma de decisiones. A continuación se llevará a cabo en la sección siguiente, el desarrollo del tópico correspondiente a memoria organizacional.

2.4 Memoria Organizacional

2.4.1 Conceptos Generales

En este apartado, se describirá en breves palabras los conceptos generales de memoria organización, con el objetivo de tener una visión general de la misma.

Al momento de definir la memoria organizacional, podemos establecer que la misma es considerada como un sistema donde se puede almacenar las experiencias vividas, percibidas, recuperar luego las mismas para luego usarlas en un tiempo futuro [87].

Cuando se habla de memoria organizacional, se considera que su definición está asociada a lo que se conoce como aprendizaje organizacional. En este sentido, la memoria organizacional y el aprendizaje organizacional son procesos sociales que ayudan a la sinergia organizacional a través de la integración de recursos tecnológicos e intercambio de conocimiento [88].

Teniendo en cuenta que para que una organización pueda aprender de las experiencias previas, las mismas deben estar disponibles para cada individuo en la organización [89]. De este modo, se considera que la memoria organizacional o memoria organizativa, no solo es considerada como una forma de retener conocimiento sino que también es posible compartir el mismo al resto de los miembros de una organización.

A continuación, en la sub sección siguiente se detallarán algunos modelos de memoria organizacional.

2.4.2 Clasificación de Modelos de Memoria Organizacional

De acuerdo a los diferentes modelos de memoria organizacional a analizar, podemos organizar la clasificación de los mismos en las siguientes categorías:

- Modelo basado en niveles de abstracción,
- Modelo basado en información,
- Modelo basado en dimensiones y
- Modelo ampliado.

A continuación, se describirán en forma general cada uno de ellos, considerando los aspectos más importantes de los mismos.

2.4.2.1 Modelo basado en niveles de abstracción

El presente modelo, de acuerdo con [90], presenta una memoria organizacional, la cual se encuentra establecida en diferentes niveles de abstracción.

En este modelo, el autor, pretende representar a cada uno de las personas involucradas en la organización, junto con los procesos (sus entradas y salidas), y las relaciones que existen entre ellos [90]. La idea principal es definir los stakeholders que crean fuentes (documentos, notas, etc) y desempeñan roles diferentes para la creación, uso y mantenimiento de varios objetos. Los objetos son creados por diferentes tareas en la organización, como ser objetos físicos, por ejemplo, documentos.

En la Figura 7 se puede ver representado el modelo de memoria organizacional basado en diferentes niveles de abstracción, de acuerdo a lo establecido por el autor.

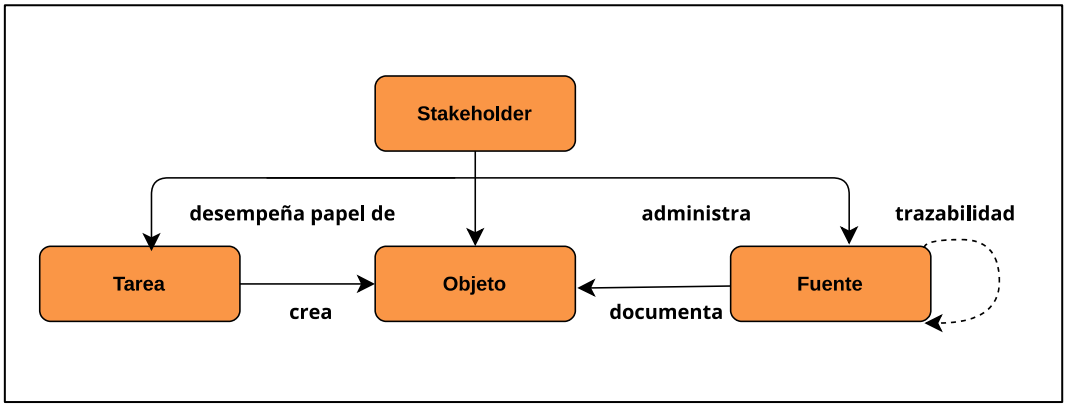

Figura 7 Modelo de Memoria Organizacional basado en niveles de abstracción [90]

Como se puede observar en la Figura 7, el autor plantea en este modelo, que la trazabilidad a través de diversos objetos, se encuentra representados a través de las trazas a los distintos enlaces establecidos.

2.4.2.2 Modelo basado en información

En este tipo de modelo, el autor [91], plantea que el mismo se encuentra basado en tres ontologías las cuales son utilizadas para describir la información y el conocimiento.

Tal como se puede observar en la Figura 8, la idea principal es que cada uno de los elementos se encuentran descriptos por un conjunto de conceptos. Dicho modelo, describe fuentes diferentes de información, considerando su estructura, formato y propiedades de acceso.

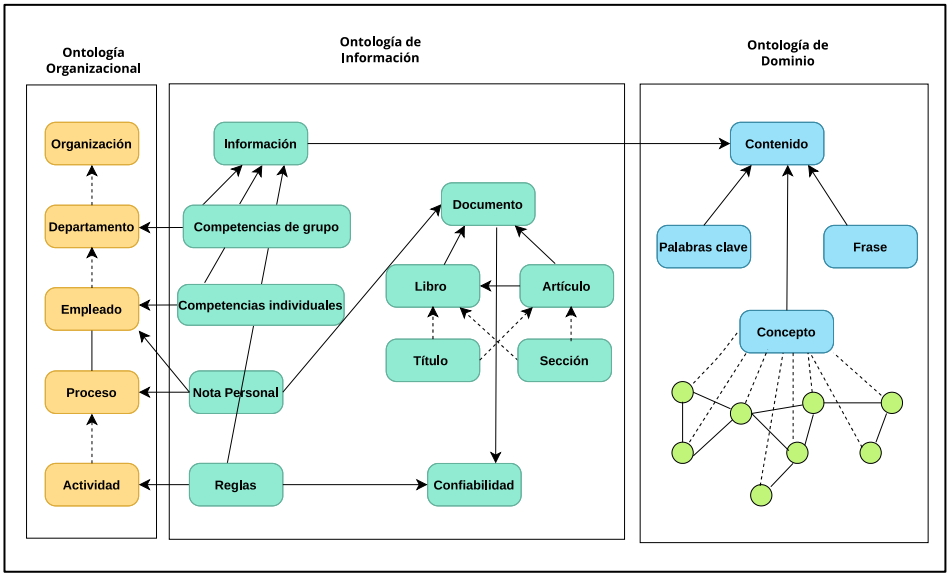

Figura 8 Modelo de Memoria Organizacional basado en información [91]

El vocabulario utilizado es provisto por la ontología de información. De este modo la ontología cubre los aspectos de las fuentes de información y de conocimiento. Por otro lado, también ofrece los enlaces a la ontología de dominio y a la ontología de organización. Cabe aclarar, que el autor en este tipo de modelo, indica que para modelar el contexto de información, utiliza la ontología de organización. Modelar el contexto tiene como tarea principal describir el contexto.

2.4.2.3 Modelo basado en dimensiones

En este tipo de modelo, los autores [92], proponen un modelo basado en tres diferentes dimensiones de memoria organizacional: a) retenedores de conocimiento; b) conocimiento en sí; y c) meta conocimiento.

Como se puede visualizar en la Figura 9, los retenedores de conocimiento se caracterizan por el meta concomiendo cognitivo y descriptivo. Dichos retenedores, poseen conocimiento de los conceptos de la organización.

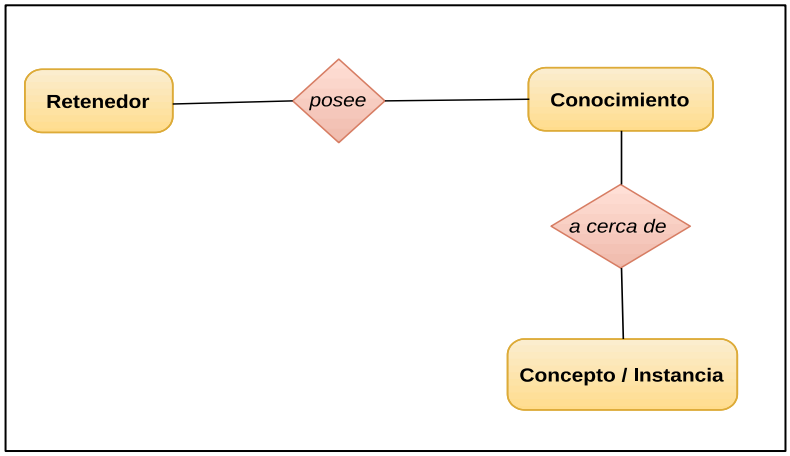

Figura 9 Modelo de Memoria Organizacional basado en dimensiones [92]

2.4.2.4 Modelo ampliado

El modelo anterior, especificado por [92]; sufre una serie de modificaciones donde el autor [93], indica la incorporación de entidades en el mismo.

La ampliación consiste en la incorporación de entidades donde los retenedores de conocimiento están relacionados a algún tipo de almacenamiento. Se crea la entidad denominada por el autor como "dominio del conocimiento", donde se pueden representar conceptos, añadiendo además sinónimos y definiciones de conceptos.

La Figura 10 muestra el modelo con su respectiva ampliación, así como también las relaciones establecidas entre cada uno de los componentes definidos, de acuerdo a lo establecido por el autor.

La idea de ampliar el modelo, consiste básicamente en poder mejorar la utilización del mismo en lo que respecta al análisis del contexto.

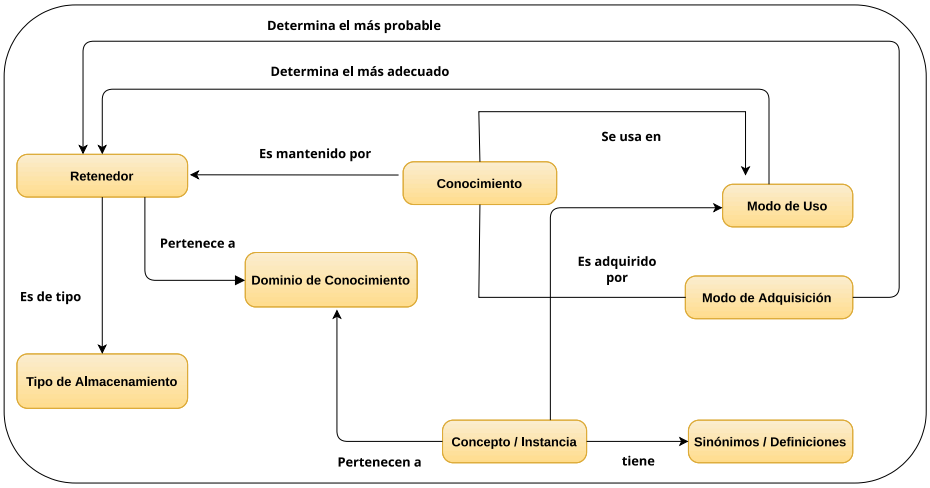

Figura 10 Modelo de Memoria Organizacional Ampliado [93]

Podemos observar que la entidad conocimiento fue vinculada a las entidades modo de uso y modo de adquisición. Esta incorporación permitirá incorporar al modelo información asociada al contexto.

En relación a los modelos de memoria organizacional antes descriptos, podemos evaluar que por sí solo el conocimiento no genera valor, pero la aplicación de dicho conocimiento crea las ventajas competitivas de una organización generando valor. Es decir, que la tecnología junto con las herramientas, los sistemas expertos y sistemas de toma de decisiones permiten brindar apoyo para diseñar y aplicar el conocimiento.

Ahora bien, el procesamiento de flujo de datos, es considerado barato al momento de utilizar el mismo, ello se debe justamente a que la tecnología en forma conjunta con la economía de escala, avanzan y evolucionan. Es más, los repositorios de Big data y los flujos de datos se alimentan permitiendo la toma de decisiones basada en datos como un aspecto natural en las diferentes organizaciones, por ejemplo los gobiernos [94].

Por otro lado, la Arquitectura de Procesamiento basada en Metadatos de Medición (en inglés PAbMM) [95], es un motor de flujo de datos orientado a proyectos de

medición, que apoya la toma de decisiones a través de una memoria organizacional. Para poder detectar situaciones tipificadas y apoyar la toma de decisiones, la arquitectura de procesamiento utiliza la memoria organizacional. Dicha memoria, se encuentra integrada por datos históricos, definiciones de los proyectos de M&E y además por las experiencias y conocimiento previos proveniente de los expertos en cada proyecto de medición. La experiencia y los conocimientos previos de los expertos son útiles para llevar a cabo las recomendaciones de las líneas de actuación relacionadas con las situaciones tipificadas.

Sin embargo, puede suceder que una situación tipificada no posea experiencia previa (o que sea una situación nueva) y en ese caso, no se dispondría de recomendaciones. Por ello se definieron coeficientes estructurales y de comportamiento a partir de la definición del proyecto [16], con el objetivo de evitar dicha situación. En este sentido, el coeficiente estructural básicamente busca atributos que caractericen a las entidades analizada, y para ello, un ordenamiento inicial permite encontrar las entidades similares y reutilizar su experiencia. Por otro lado, y para mejorar la precisión, el coeficiente de comportamiento analiza los atributos comunes de las entidades basándose en la correlación entre la distribución de los datos. De este modo, una puntuación considerando la estructura y su comportamiento asociado, permite buscar entidades similares para reutilizar su experiencia y conocimientos.

En [96], Zhou, Wun y otros proponen un modelo basado en la navegación del usuario y el aprendizaje de las preferencias de navegación, los diferentes tipos de relación social y el mapeo para homogeneizar la estructura de los datos manteniendo sus propiedades bajo un modelo supervisado. La idea en relación con el modelo es que tratan de que el siguiente enlace relacionado con el comportamiento de navegación de un determinado usuario, se pueda predecir. En la presente tesis, la idea está orientada a buscar proyectos similares de M&E cuando no se dispone de experiencia previa. En este sentido, y con el objetivo de mantener en memoria las recomendaciones que se consideren más cercanas en términos de los proyectos de medición, es que se plantea el uso del coeficiente estructural en la arquitectura para determinar aquellos proyectos que sean similares en función de su definición. Por otro lado, se utiliza en forma conjunta el coeficiente de comportamiento con el estructural para calcular una puntuación similar entre los proyectos de M&E. Así, mediante el uso de los coeficientes estructurales y de comportamiento, los proyectos más relacionados son guardados en la memoria a fin de aportar recomendaciones cuando alguna alarma en tiempo real sea recibida por el responsable de tomar decisiones [97].

Es un buen punto de partida, indicar las diferencias que existen entre la memoria organizacional y la memoria institucional. En este sentido, en [98], los autores plantean a través de la aplicación de una Revisión Sistemática de la Literatura, que la memoria institucional está relacionada con el valor social del grupo, tiene dificultad para ser práctica, realista y objetiva. Está en el conjunto instituido, no es pragmática y es un fenómeno que trata las relaciones de poder. Por otro lado, la memoria organizacional,

es práctica, objetiva, se encuentra centrada en acciones concretas y en la productividad. Sin embargo, así como se plantean las diferencias más significativas, los autores indican que ambos procesos de memoria son cíclicos y se encuentran en constante construcción.

Por otro lado, se plantea en [68], una ontología de memoria organizativa basada en casos con el objetivo de contribuir al aprendizaje, razonamiento y resolución de problemas que contribuyan a la toma de decisiones. En dicho trabajo, los autores establecen que el objetivo de la memoria organizacional es servir de base para el intercambio de conocimiento organizacional en una arquitectura de procesamiento de datos , basada en la medición y evaluación. La arquitectura que proponen utiliza un repositorio de big data para que los datos estén a disposición para su consumo y de este modo poder gestionar la memoria organizativa.

En lo que respecta a las ontologías para modelar una memoria organizativa, podemos mencionar que existen numerosos modelos y herramientas que permiten la gestión del conocimiento. En [99], los autores plantean una ontología de memoria organizativa basada en casos para contribuir al diseño de una memoria organizativa que permita contribuir y dar apoyo a una mejor toma de decisiones.

2.5 Conclusiones Generales del Capítulo

Los temas presentados en el presente capítulo, abordan las bases generales del contexto para comprender el mismo como así también poder especificar las principales contribuciones de este trabajo de tesis.

En primer lugar se planteó la utilización de la metodología Estudio de Mapeo Sistemático de la Literatura (SMS) a fin de poder identificar y encontrar toda la evidencia posible asociada a los conceptos de sistemas de recomendación, marcos de medición y evaluación, medidas de similitud y memoria organizacional.

A partir del objetivo general del trabajo de tesis, donde se planteó *"Desarrollar una estrategia de recomendación en memoria, basada en entidades bajo monitoreo semánticamente similares, para mejorar la precisión y reutilización de conocimiento y/o experiencia previa ante situaciones nuevas y-o no tipificadas, en los cuales una decisión requiera de cursos de acción como soporte",* se hizo énfasis en la necesidad de integrar todos los conceptos descriptos en el capítulo.

En primer lugar se describieron trabajos asociados a los sistemas de recomendación, en donde a partir de la aplicación del SMS, se analizaron los distintos tipos, se evaluaron las áreas de aplicación, técnicas y métodos junto con las ventajas que presenta cada uno de ellos. Del resultado del SMS, se encontraron que existen diferentes aplicaciones de los sensores IoC y los sistemas de recomendación en áreas como la agricultura, la seguridad, la salud, entre otras. La aplicabilidad de los sistemas de recomendación en tiempo real y en la toma de decisiones sin dudas, dependen del contexto en el cual se encuentran inmersos.

Seguidamente, se describieron diferentes marcos de medición y evaluación, junto con las aplicaciones de los mismos en distintas áreas. Analizando la eficacia de cada uno de ellos, así como también la ventaja de utilizar una estructura que permita brindar una base para los proyectos de medición y evaluación. Se detalló en particular los distintos componentes del marco conceptual C-INCAMI como una forma de introducir al marco general que forma parte de la propuesta de tesis, para luego hacer hincapié en la extensión del mismo. Se encontró que existen marcos que permiten la toma de decisiones multidimensional, como así también existen marcos que posibilitan el monitoreo en entornos de virtualización y otros que posibilitan el medir y evaluar el rendimiento de los sistemas de transporte.

En este capítulo también fueron presentados los conceptos y clasificaciones de similitud semántica a partir de la ejecución del SMS correspondiente. Se detectaron que existen procesos organizados para poder medir y comparar ontologías, así como también la aplicación de similitud semántica relacionada con la entidad bajo monitoreo. Por otro lado, se encontró una nueva técnica de cálculo de la similitud semántica entre dos entidades, donde la técnica se basa en el conocimiento de una ontología o taxonomía.

En la última sección del capítulo, se describió el concepto de memoria organizacional, donde se detallaron los diferentes modelos asociados a la misma, planteando las distintas aplicaciones de cada uno de ellos.

Finalmente y de acuerdo al objetivo del presente trabajo de tesis, la idea consiste en poder desarrollar una estrategia que permita buscar en memoria y determinar la similitud semántica entre entidades bajo monitoreo en proyectos de M&E basados en el marco conceptual C-INCAMI, para de esta forma poder brindar recomendaciones en tiempo real cuando no exista experiencia o conocimiento previo sobre una entidad. Con ello se pretende mejorar la precisión de las recomendaciones a brindar en determinadas situaciones, reutilizar todo el conocimiento o experiencia previa ante determinadas situaciones (nuevas o no) basada en la similitud semántica de las entidades que se encuentran bajo monitoreo. Los detalles de la solución al problema planteado se describirán en los capítulos siguientes.

Capítulo 3	Formalización del Proyecto de Medición y Evaluación

Capítulo 3. Formalización del Proyecto de Medición y Evaluación

Introducción

El proceso de medición puede definirse como el proceso en el cual un objeto o sujeto bajo análisis necesita ser cuantificado a través de uno o más atributos característicos [100]. La medición consiste en un esquema de cuantificación donde se intenta contrastar el valor obtenido contra un patrón de referencia. Por ejemplo, la altura de una persona requiere contar con un patrón de referencia como el metro, el peso de un vehículo requiere un contraste contra el kilogramo, y así sucesivamente. Estos ejemplos permiten observar dos desafíos diferentes. Por un lado, la necesidad de cuantificar y obtener un cierto número vinculado a una característica de un objeto, sujeto, o concepto. Por otro lado, es fundamental contar con un patrón de referencia a los efectos de la comparación.

Estos desafíos nos llevan a incógnitas esenciales relacionadas a la necesidad de medir y de comparar los resultados. En tal sentido, las personas tienden a analizar y comparar cada comportamiento, fenómeno, objeto, concepto, y/o sujeto de su entorno. La evolución del ser humano requirió una profunda comprensión de su entorno como de sus interacciones, por ejemplo, poder discernir entre frío y calor, caro y costoso, alto o bajo, pesado o liviano, entre otros. De hecho, uno de los primeros conceptos desarrollados por los humanos fue el número a los efectos de contabilizar objetos en un contexto (por ejemplo, alimentos) [101]. Esto permitió la contabilización y también establecer diferencias a partir de ellas basado en la comparación. Pero el hecho de que todos puedan obtener sus propios números requirió de un patrón uniforme y común útil para establecer analogías entre los conceptos [102]. De este modo, surgieron diferentes sistemas como el métrico.

En la actualidad, el proceso de medición se aplica en múltiples áreas, y puede ser caracterizado como transversal y no exclusivo debido a su amplio alcance. Sin embargo, es importante mencionar que el modo en que la medición se obtiene es crítico a los efectos de contrastarlo respecto de un patrón de referencia. Es decir, el punto crítico de la medición es la comparación, por lo cual, se supone que las magnitudes son comparables. Por ejemplo, si se desea medir la temperatura corporal de una persona, podría emplearse un dispositivo bajo la región axilar, o alternativamente, en la zona interior del oído. En ambos casos se podría obtener un número, aunque para el mismo instante de tiempo pueden variar. Más aún, dicha variación no se debería al instrumento utilizado sino a la zona donde se obtuvo la medida. En tal caso, no se podría decir que la temperatura corporal subió o bajó dado que se intentarían comparar valores provenientes de zonas corporales complementarias, pero no equivalentes [103].

Por tal, es posible apreciar que la intención de medir para cuantificar un concepto conlleva a un conjunto de acuerdos involucrados. Es decir, se requiere identificar la entidad a ser monitoreada como la característica en particular a cuantificar. Debe

detallarse el modo en que el valor numérico se obtiene desde la característica mediante un instrumento y técnica dada. Además, se debiera especificar cómo garantizar la comparabilidad de las medidas sobre el tiempo. Por ejemplo, si la técnica de medición en un tiempo ´t+1´cambiare, debiera garantizarse que los valores respecto al concepto son comparables (independientemente de aspectos de precisión), de lo contrario, se estaría en riesgo de tornar en incomparable la serie histórica de medidas o limitar su comparabilidad.

Los cambios globales, el impacto tecnológico, y el dinamismo de las economías requieren procesos de mediciones ágiles y adaptables, con posibilidades de procesamiento de datos en tiempo real [104]. Por tal, la idea es aproximar una alternativa de medición ágil, extensible, confiable, escalable, y estable como sea posible. No obstante, existen un gran número de conceptos (como las relaciones entre ellos) que requieren ser acordados para automatizar un proceso de medición (por ejemplo, la idea de métrica, medida, medición, escala, unidad, método, entre otros). Por tal motivo, antes de intentar automatizar o implementar un proceso de medición, es crítico acordar y compartir los conceptos y sus relaciones a lo largo del proceso de medición e interesados para garantizar la confiabilidad del sistema [105].

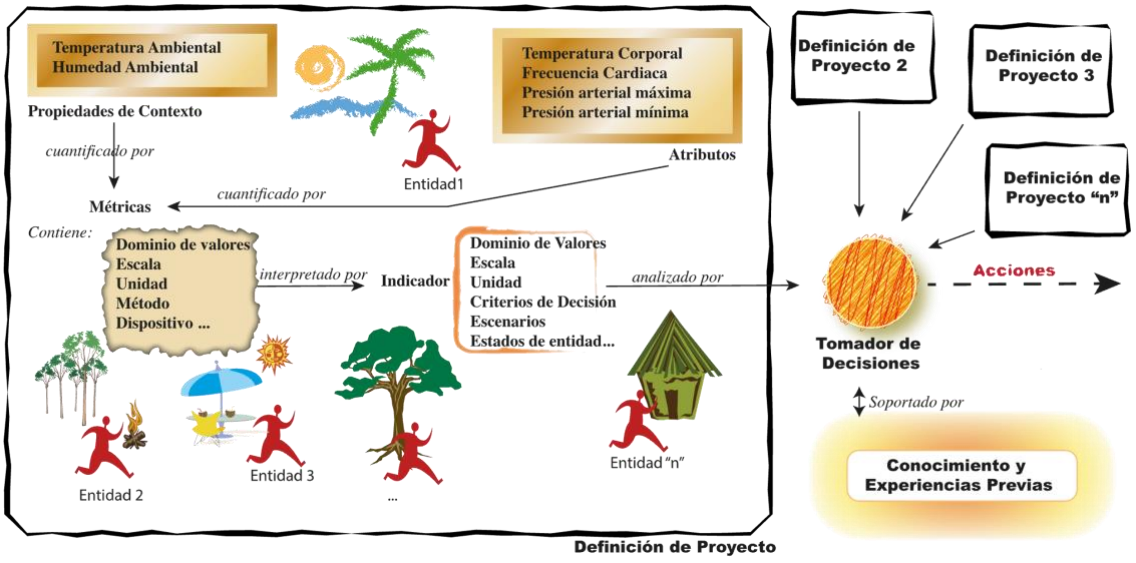

Figura 11 Perspectiva Global de los Conceptos Involucrados en un Proceso de Medición

Como se puede apreciar en la Figura 11, el concepto a ser monitoreado (Entidad) necesita ser descrito por un conjunto de atributos (por ejemplo, frecuencia cardiaca). Esto implica una representación discreta de los aspectos a ser analizados para la entidad. Además, la entidad analizada se encuentra inmersa en un contexto que es descrito a través de un conjunto de propiedades contextuales (por ejemplo, temperatura ambiental). Esto implica a su vez que el contexto con el que interactúa la entidad es

interpretado discretamente. Tanto los atributos como las propiedades de contexto son estudiados en forma conjunta para analizar la incidencia mutua.

De este modo, cada atributo o propiedad de contexto es cuantificado por una métrica. Cada métrica contiene un dominio de valores esperados, una escala, unidad, método a emplear para obtener el valor numérico, un instrumento, entre otros aspectos. Esta definición es particularmente importante porque permite saber si dos medidas son comparables. Por ejemplo, si se emplearon métodos compatibles o equivalentes. Para el ejemplo anteriormente introducido de la temperatura corporal, las medidas de los métodos nos serían comparables directamente porque la temperatura intra auditiva versus la axilar presentan una diferencia de alrededor de 0.5ºC [103].

Sin embargo, las métricas y medidas proveen un valor numérico (por ejemplo, 37.8ºC para la temperatura corporal), pero nada dice respecto de cómo interpretar el mismo. En otras palabras, la medida no aporta a los efectos de poder concluir sobre si 37.8ºC es normal o no en determinadas situaciones o condiciones. En este punto es donde el concepto de indicador toma importancia. El indicador consume una o más medidas provenientes de las métricas (incorporando los criterios de decisión basados en los estados de entidad y escenarios definidos) para analizar la magnitud en su contexto. De este modo, el indicador permite interpretar el valor y poder concluir sobre si es normal o no en un escenario/estado dado. Así, la figura del tomador de decisiones emplea el indicador como medio de interpretación de las medidas, mientras que capitaliza el conocimiento y experiencia previa para proveer recomendaciones (o cursos de acción) basado en la interpretación del indicador. Por ejemplo, el indicador (mediante sus criterios de decisión) podría indicar que 37.8ºC no es normal y la persona estaría con fiebre de acuerdo con su estado y escenario actual, a lo que el tomador de decisiones (empleando la experiencia previa) podría recomendar acciones orientadas a controlar/disminuir la fiebre.

Esta interpretación de los conceptos requeridos respecto de la medición como de su proceso, requiere una definición y organización clara, comunicable, y compartible disponible para quien requiera utilizarla. En este sentido es donde el “framework” toma especial énfasis. El “framework” de medición debe proveer todos los términos, conceptos y relaciones disponibles que sean esenciales para implementar un proceso de medición repetible, consistente, y extensible. La repetibilidad permite aplicar en reiteradas oportunidades el mismo proceso y obtener medidas comparables. La consistencia implica que cualquier cambio a la definición del proyecto no afectaría negativamente la comparación con medidas históricas. La extensibilidad refiere a la posibilidad de actualizar el proceso de medición ante nuevos requerimientos de forma transparente.

De este modo, el presente capítulo se organiza alrededor de cuatro secciones principales. La primera sección discute los conceptos, términos, y las relaciones entre ellos a partir de un abordaje ontológico de C-INCAMI. Allí, se discuten nuevos conceptos

que han sido incorporados al marco de medición original para soportar conceptos tales como estados de entidad, escenarios contextuales, criterios de decisión basados en escenarios y estados de entidad, entre otros. La segunda sección introduce la evolución de la estrategia GOCAME a GOCAME-ESVI para soportar los nuevos conceptos ontológicos e incorporar guías de visualización para los indicadores. La tercera sección discute un nuevo esquema de organización y formato del proyecto de medición (BriefPD) que permite optimizar la comunicación de múltiples proyectos de medición en forma autocontenida. Finalmente, la sección 4 concluye sobre la integración del marco de medición C-INCAMI extendido, el rol de GOCAME-ESVI y BriefPD en el proceso de recolección de datos distribuidos en entornos heterogéneos.

El capítulo se soporta en las siguientes publicaciones efectuadas a lo largo del proceso de investigación de acuerdo con cada sección mencionada:

- Extensión del Marco de Medición (ECINCAMI)
    - Diván, M & Sánchez Reynoso, M (2020) "**A Real-Time Entity Monitoring based on States and Scenarios**" CLEI Electronic Journal. ISSN 0717-5000. Vol. 23 (1). Pp 2-1:2-25. https://doi.org/10.19153/cleiej.23.1.2

    - Diván, M and Sánchez Reynoso (2019) "**Extending the Data Stream Processing Strategy to Scenario Analysis**". In proceedings of International Conference on Innovations in Computer Science and Engineering (ICICSE).26-28 June of 2019. Miri, Sarawak, Malaysia. International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE). World Academy of Research in Science and Engineering (Publisher). 8(1.4):1-8. ISSN 2278-3091. https://doi.org/10.30534/ijatcse/2019/0181.42019

    - Diván, M & Diván, M (2019) "**Incorporating Scenarios and States Definitions on Real-Time Entity Monitoring in PAbMM**". XLV Latin American Computing Conference, {CLEI} 2019, Panama, September 30 - October 4, 2019. IEEE. https://doi.org/10.1109/CLEI47609.2019.235072

- El proceso de medición y Guías de Visualización
    - Sánchez Reynoso, M &Diván, (2020) ***"Applying Data Visualization Guideline on Forest Fires in Argentina".*** 10th International Conference - **CONFLUENCE' 2020**. Department of Computer Science and Engineering. Amity University. Uttar Pradesh, India. January 29 – 31 of 2020. https://doi.org/10.1109/Confluence47617.2020.9058174

    - Sánchez Reynoso, M &Diván, (2019) ***"Contributions to the Communication of the Official Advertising's Distribution in Argentina".*** 4th International Conference on Information Systems and Computer Networks (**ISCON**). Department of Computer Engineering & Applications,

GLA University, Mathura (UP), India. November 21 – 22 of 2019.https://www.gla.ac.in/iscon2019/index.html
DOI: 10.1109/ISCON47742.2019.9036298

- Diván, M & Sánchez Reynoso, M (2018) "**The Real-Time Measurement and Evaluation as System Reliability Driver**". Book Chapter in "System Reliability Management: Solutions and Technologies". Anand, A & Ram, M (Eds.). CRC Press, Taylor & Francis Group. Pp. 161-188. https://doi.org/10.1201/9781351117661-11

- Comunicabilidad del proyecto de medición
  - Diván, M. Sánchez-Reynoso, M. Gonnet, S. (2022) **"Measurement Project Interoperability for Real-time Data Gathering Systems**". Future Generation Computer Systems, Elsevier, ISSN 0167-739X, 129, 298-314 https://doi.org/10.1016/j.future.2021.11.031

3.1 Ontología de Medición y Evaluación: Extendiendo C-INCAMI

El marco de medición y evaluación C-INCAMI plantea una serie de conceptos, términos, y relaciones necesarias para implementar un proceso de medición y evaluación [106], [107]. Este se complementa con la estrategia GOCAME (Goal-oriented Context-aware Measurement and Evaluation)[108] que define el modo en que un proyecto de medición y evaluación es definido de acuerdo con los conceptos y términos especificados. Sin embargo, el marco no contempla la posibilidad de modelar cambios de estados en las entidades bajo monitoreo, ni tampoco el cambio de escenarios para los contextos. Estos últimos aspectos son de esencial importancia para los indicadores, quienes emplean criterios de decisión para interpretar un valor numérico [109]. Es decir, un mismo indicador puede tener interpretaciones diferentes de acuerdo con el estado actual de la entidad y su escenario. Por ejemplo, la frecuencia cardiaca de una persona podría ser 170 pulsaciones por minuto. Si se interpreta el valor en forma aislada podría indicarse que sería un valor atípico para una persona de 40 años. Sin embargo, si se indica que la persona está realizando ejercicio físico (estado de la entidad) en un ámbito cerrado como un gimnasio con determinada temperatura (situación del escenario), se interpretaría como que se encuentra en plena actividad y el valor se podría considerar normal. En otras palabras, el mismo valor numérico podría interpretarse diferente de acuerdo con el estado actual de la entidad y su escenario inmediato.

Figura 12 introduce la serie de nuevos conceptos, términos, y relaciones incorporadas y describe su articulación con los conceptos originales de C-INCAMI. Esta extensión da lugar a una versión extendida de la ontología de CINCAMI que se ha denominado ECINCAMI por Extended-CINCAMI. ECINCAMI está descrita mediante OWL (Ontology

Web Language) y se encuentra disponible bajo los términos de la licencia *Creative Commons 4.0* en *Figshare* y en el *Repositorio y Registro de Ontologías*.

Sintéticamente, ECINCAMI es una ontología de medición basada en la necesidad de información descrita por el usuario que es quien define el proyecto de medición. Esto permite identificar y describir una categoría de entidad a monitorear (ej., una persona) y caracterizarla a través de distintos atributos (ej., frecuencia cardíaca, temperatura corporal, etc.). El entorno con el que interactúa la entidad permite identificar un contexto, el cual es caracterizado mediante propiedades de contexto (ej. temperatura ambiental, humedad, etc.). Las métricas permiten cuantificar tanto los atributos de la entidad como las propiedades del contexto para obtener un valor numérico. Sin embargo, la interpretación de los valores numéricos es realizada mediante los indicadores. Un estado de entidad [110] es una configuración observable particular de sus atributos característicos que permite un análisis o interpretación diferencial. Por ejemplo, de acuerdo con la frecuencia cardíaca y presión arterial, una entidad podría estar en uno de los siguientes estados: caminando, corriendo, o descansando. Un escenario [111] describe una configuración observable de las propiedades de contexto que permiten analizar o interpretar el entorno. Por ejemplo, basado en la temperatura ambiental y humedad, el contexto puede corresponder con escenarios extremo, normal, o ideal para la práctica de una actividad física de una persona de 40 años. Los indicadores interpretan los valores numéricos de las métricas empleando criterios de decisión basados en la situación del escenario y estado de la entidad. Tal interpretación, por ejemplo, define si hay riesgo o no para la salud de una persona de 40 años que está corriendo (estado) en un escenario extremo (escenario). Como consecuencia de la interpretación del indicador, un conjunto de recomendaciones o acciones podrían derivarse (ej. Notificar a la persona que está en riesgo e informar a un servicio de salud próximo a ella)[112].

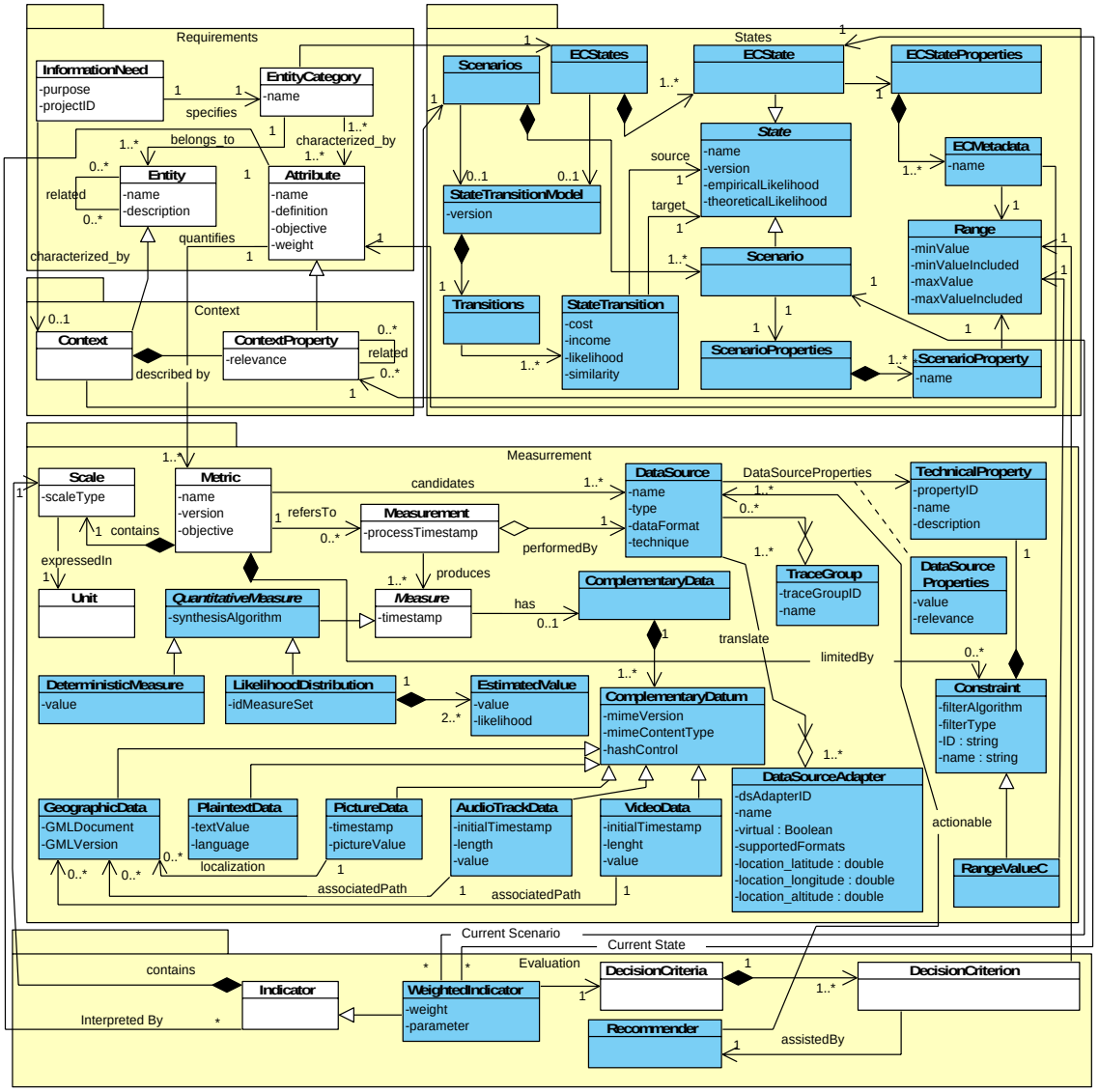

Figura 12 Principales Conceptos del Marco Extendido ECINCAMI

La figura anterior describe los conceptos principales de la ontología ECINCAMI organizados en cinco paquetes: requerimientos, contexto, medición, estados, y evaluación. Los nuevos conceptos respecto del marco original CINCAMI se indican con fondo celeste, mientras que los originales se representan con fondo blanco. Estos nuevos conceptos permiten una nueva organización jerárquica de la definición del proyecto de medición que se abordará en la sección 3.3.

El paquete *requerimientos* describe el objetivo y la entidad para un proyecto de medición. La clase *EntityCategory* describe el concepto a ser analizado (ej. un paciente ambulatorio), mientras que la clase *Entity* representa cada instancia. La clase *InformationNeed* describe el objetivo del proyecto para la categoría de entidad bajo monitoreo. Cada clase *EntityCategory* es caracterizada por un conjunto de atributos cuantificables (ej. la frecuencia cardíaca). La propiedad *weight* en la clase *Attribute* señala la importancia relativa de cada uno respecto del proyecto para la categoría de

entidad. Es decir, representa un acercamiento multidimensional y discreto para la medición de la categoría de entidad.

La clase *Context* es un tipo especial de entidad que representa el ambiente donde una entidad desarrolla una actividad (Ver el paquete *Context*). El contexto es descrito por medio de un conjunto de instancias de ContextProperty (es decir, propiedades de contexto) que son una especialización de la clase *Attribute*. La propiedad *relevance* indica la importancia relativa de cada propiedad de contexto en un entorno dado. Por ejemplo, es un modo de señalar si la temperatura o la humedad tienen mayor prioridad en el análisis cuando una persona desarrolla actividad física. Tanto los atributos como las propiedades de contexto son cuantificadas mediante métricas.

El paquete measurement contiene los elementos para cuantificar los atributos de una entidad o las propiedades del contexto. La métrica es representada a través de la clase Metric y es responsable por obtener una representación numérica. Tal número tendrá asociada una escala y unidad (Ver las clases Unit y Scale en la Figura 12). Cada vez que el proceso de medición es ejecutado (Ver clase Measurement) produce un conjunto de medidas (Ver clase Measure). La clase QuantitativeMeasure es una especialización de la clase Measure. Por un lado, una fuente de datos (o dispositivo) puede producir un valor numérico simple (Ver clase DeterministicMeasure). Por otro lado, es posible estimar un valor a través de un conjunto de pares (probabilidad, valor), el cual es representado mediante las clases LikelihoodDistribution y EstimatedValue. A los efectos de obtener un valor numérico simple desde una distribución de probabilidad, la clase QuantitativeMeasure incorpora la propiedad synthesisAlgorithm.

Opcionalmente, una medida podría tener asociado un conjunto de datos complementarios, tales como una ubicación, texto plano (ej. con datos de un registro de transacciones u operaciones), imagen, pista de audio, o secuencia de video (Ver las clases GeographicData, PlaintextData, PictureData, AudioTrackData, y VideoData). Todos ellos son capturados al mismo instante en que la medida es obtenida (Ver la clase ComplementaryData).

Cada medida es obtenida siguiendo la definición de la métrica. Esta se implementa por una fuente de datos o sensor (Ver la clase DataSource en el paquete Measurement), satisfaciendo un conjunto de restricciones (Ver las clases TechnicalProperty, Constraint, y DataSourceProperties). La clase TraceGroup representa un conjunto de sensores monitoreando un concepto similar. La clase DataSourceAdapter representa el dispositivo para el cual los sensores (o fuentes de datos) están directamente conectados.

El paquete *States* describe el conjunto de estados de entidad y escenarios. Cada atributo de entidad es asociado con una clase *ECMetadata*. Así, un conjunto de instancias *ECMetadata* (Ver la clase *ECStateProperties*) describen un estado de entidad (Ver la clase *ECState*). Análogamente, cada propiedad de contexto está relacionada a una clase *ScenarioProperty*. Un conjunto de instancias *ScenarioProperty* (Ver la clase

ScenarioProperties) caracteriza un escenario (Ver la clase *Scenario*). Las clases *ECState* y *Scenario* especializan la clase *State*, la cual incorpora la probabilidad de ocurrencia empírica y teórica. La transición entre estados es modelada a través de la clase *StateTransition*. Un conjunto de instancias de la clase *StateTransition* componen las transiciones (Ver la clase *Transitions*) y transitivamente el modelo de transición de estados (Ver la clase *StateTransitionModel*). Así, la clase *Scenarios* está integrada por un conjunto de instancias de Scenario (escenarios) y también de su modelo de transición de estado asociado. Análogamente, la clase *ECStates* se compone de un conjunto de instancias ECState (estados de entidad) y de su correspondiente modelo de transición de estado.

El paquete *Evaluation* describe los criterios de decisión definidos por expertos para interpretar los atributos. La interpretación de cada medida es realizada mediante la clase *Indicator*. Dado que los indicadores pueden tener una importancia relativa diferente, una nueva clase *WeightedIndicator* especializa la clase *Indicator*. Cada instancia de la clase *WeightedIndicator* intepreta medidas desde los atributos, utilizando un conjunto de instancias de la clase *DecisionCriterion*. La clase *DecisionCriteria* se compone de un conjunto de criterios de decisión. Cada criterio de decisión se compone de un escenario y estado de entidad. De este modo, la interpretación se circunscribe a dicha configuración de escenario y estado de entidad. Por ejemplo, si se desea interpretar si la frecuencia cardiaca de 170 pulsaciones por minuto es normal, debiera realizarse a sabiendas de la combinación de escenario y estado de la entidad (ej. la persona se encuentra descansando en un escenario extremo). Basado en ello, el indicador podría concluir que 170 es una elevada frecuencia para estar descansando y posiblemente habría sido afectada por la configuración del escenario. De este modo, con base en el criterio de decisión y su interpretación soportada en la combinación escenario-estado, es posible asociar un recomendador que provea cursos de acción que ayuden a mitigar riesgos (Ver la clase *Recommender*).

De este modo, las actualizaciones de la ontología de medición han permitido ahora:

1. Gestionar datos deterministas y distribuciones (valor, probabilidad) lo que permite aproximar una medida incluso en entornos no deterministas o con limitaciones de instrumental,
2. Incorporar datos complementarios a la medición (Video, audio, texto, ubicaciones, o imagen) lo que favorece la contextualización del valor numérico y su seguimiento mediante su ubicación,
3. Gestionar estados de entidad como configuraciones observables de las condiciones de una entidad definida a partir de sus propios atributos. Esto permite una interpretación discreta y ad-hoc de los estados de entidad como el director de proyecto de medición requiera,
4. Gestionar escenarios como configuraciones observables de las condiciones del contexto definidas a partir de sus propiedades contextuales. Esto

permite una interpretación discreta y ad-hoc de los escenarios del contexto o ambiente tan detallado como el director de proyecto de medición requiera,

5. Estimar probabilidades empíricas de transición de estados y escenarios mediante los modelos de transición. Esto permite aproximar los estados o escenarios a los que se podría transitar una entidad o contexto respectivamente, utilizando su historia inmediata.
6. Gestionar fuentes de datos heterogéneas y sus restricciones para guiar el emparejamiento de acuerdo con los requerimientos de medición (ejemplo, exactitud y precisión),
7. Definir grupos de seguimiento entre fuentes de datos para permitir múltiples puntos de vistas para monitorear un concepto mediante diferentes sensores,
8. Gestionar el rol del adaptador de fuentes de datos (o puente semántico) responsable por embeber metadatos junto a las medidas, lo que permite un procesamiento guiado por metadatos (es decir, por el significado de los datos),
9. Gestionar indicadores con sus criterios de decisión basados en escenarios y estados de entidad para lograr mayor exactitud al momento de interpretar una medida.
10. Ponderar los indicadores como modo de priorización de estos basado en estados y escenarios,
11. Asociar la figura del recomendador con los criterios de decisión de modo de poder proveer cursos de acción o recomendaciones a partir de un criterio dado.

Introducidos los nuevos conceptos y su relación con el marco CINCAMI original, la siguiente sección describe la actualización de la estrategia de definición de proyectos de medición. De este modo, se consideran los nuevos términos y conceptos ontológicos incorporados junto con las guías de visualización para comunicar los indicadores.

3.2 GOCAME-ESVI

La estrategia GOCAME (Goal-oriented Context-aware Measurement and Evaluation) incorpora una interesante perspectiva de procesos para definir el proyecto de medición y evaluación basado en el marco C-INCAMI original. Sin embargo, debido a la extensión ontológica del marco C-INCAMI (en adelante, ECINCAMI por Extended-CINCAMI) GOCAME no cubre todos los conceptos necesarios para la definición de un proyecto de medición y evaluación en este nuevo escenario.

De esto modo, se ha extendido el marco GOCAME incorporando:

1) Los nuevos conceptos vertidos en ECINCAMI,

2) Las guías de visualización de acuerdo con Munzner [113] para articular la definición de indicadores respecto de su comunicación gráfica,

Así, el nuevo marco extendido se denomina GOCAME-ESVI (Entity States, Scenarios, and Visualization). Si bien los aspectos de visualización no son incorporados dentro de la estrategia de procesamiento de flujos de datos en sí, estos son descritos en la definición de proyecto de medición para definir cómo visualizar los indicadores. De acuerdo con Munzner [113], independientemente del conjunto de pasos que integran el proceso de visualización, hay aspectos que se deben considerar a lo largo de este: Qué y Cómo. En primer lugar, es importante describir la audiencia objetivo, dominio subyacente, y las acciones necesarias para comunicar visualmente el dato (esto es el ´Qué´). Seguidamente, la codificación visual de los datos, las reglas de diseño o recomendaciones, elementos visuales soportando tareas o requerimientos, junto con el nivel de aceptabilidad de la audiencia (es decir, el ´Cómo´) [114], [115].

La Figura 13 sintetiza la idea sobre las etapas principales del proceso de visualización (en otras palabras, el ´Qué´ y ´Cómo´). Durante la etapa denominada ´Qué´ existen tres actividades a desarrollar: Análisis de los datos crudos, Análisis de datos abstractos, y Análisis de los datos a ser visualizados. El Análisis de los datos crudos consisten en identificar cada fuente de datos, su estructura y significado asociado para limitar los datos potencialmente útiles en términos de los requerimientos del usuario (es decir, asociados con el proyecto de medición y el concepto bajo monitoreo). El Análisis de los datos abstractos implica un preprocesamiento de los datos junto con su enriquecimiento para alinearlos con los requerimientos. De este modo, el Análisis de los datos a ser visualizados define qué datos y de qué modo serán visualizados.

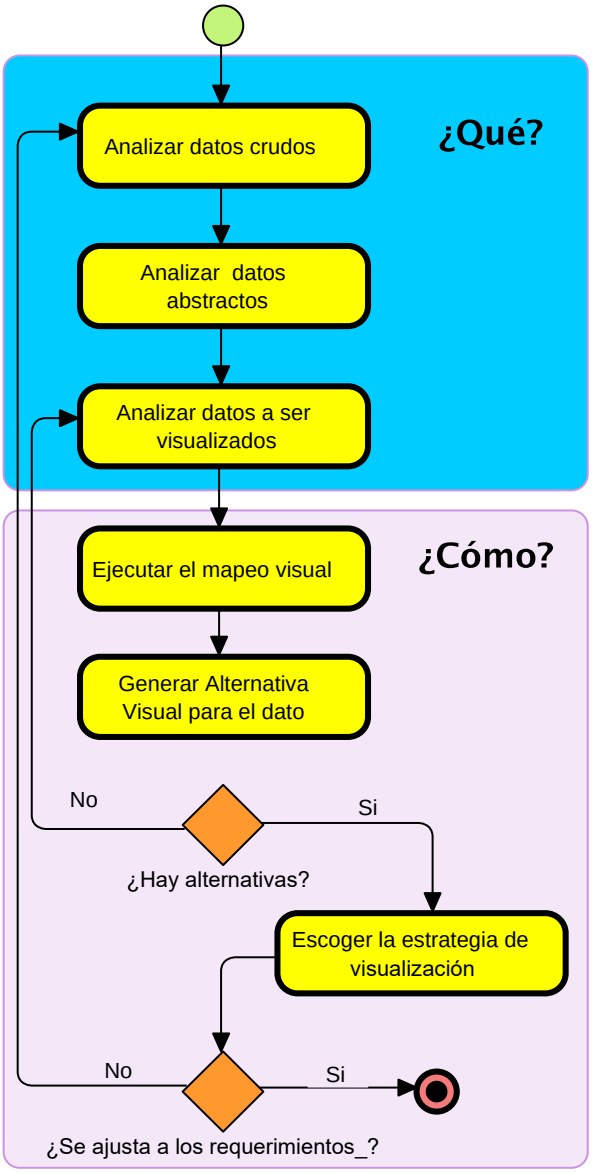

Figura 13 Tubería de Visualización expresada como un diagrama BPMN

En este sentido, es importante considerar que las variables involucradas para su visualización podrían ser categóricas, ordinales, o cuantitativas. Dependiendo de cada tipología de datos, el gráfico podría sensiblemente variar. Más aún, los datos a visualizar no necesariamente podrían asociarse con el ámbito de los datos estructurados, podrían también asociarse con datos semi-estructurados (ejemplo, registros de operaciones de un recolector de datos), o no estructurados (ejemplo, una pista de audio).

La segunda fase de la tubería de visualización se asocia con el ´Cómo´ debido a que allí el diseño de la implementación y su implementación deben llevarse a cabo. El mapeo visual de datos define el modo en que cada dato se muestra junto con la estrategia de representación de sus relaciones. Así, el diseño indica la estructura visual pertinente, los atributos a codificar, como así también la relación entre atributos y gráficos. El diseño necesita satisfacer los criterios de expresividad y efectividad, los cuales son evaluados al

momento de generar cada alternativa para saber si satisfacen o no los requerimientos del usuario.

De este modo, GOCAME-ESVI [116] incorpora los nuevos conceptos ontológicos del marco ECINCAMI junto con la tubería de visualización de Munzner.

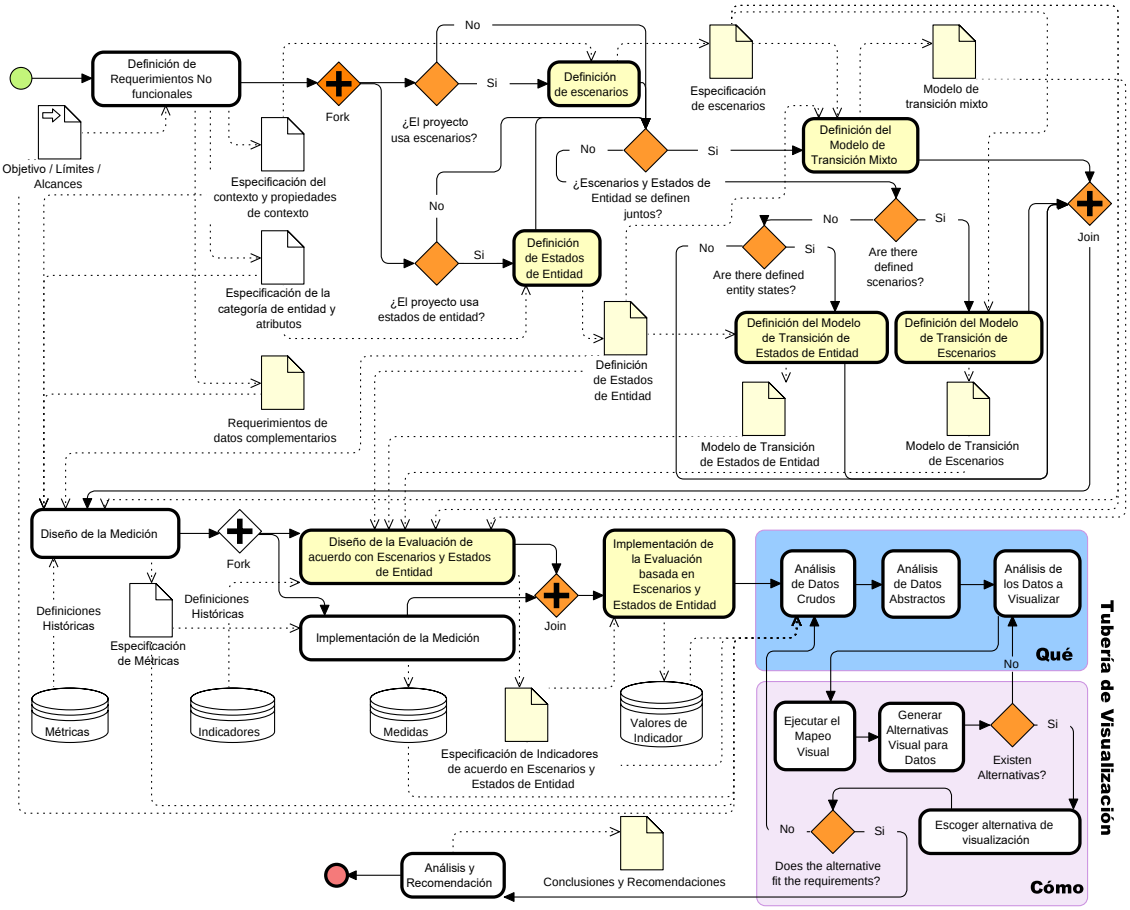

Figura 14 Una Descripción Global de GOCAME-ESVI utilizando notación BPMN

La Figura 14 introduce una descripción general de GOCAME-ESVI utilizando notación BPMN. El proceso comienza con la especificación del objetivo, límite, y alcance del proyecto de medición. Seguido, se define la categoría de entidad (ejemplo, paciente ambulatorio), los atributos a analizar (ej., frecuencia cardíaca, presión arterial, etc.), contexto (ej., aire libre, casa, etc.), las propiedades de contexto en consideración (ej., temperatura ambiental, etc.) y los datos complementarios cuando sean requeridos (ej., posición geográfica). Dependiendo del objetivo del proyecto de medición, es posible que la especificación de contexto, escenarios, o estados de entidad no sean requeridos (es decir, son opcionales). Los escenarios se definen a partir de una combinación de propiedades de contexto con intervalos predefinidos de valores. Análogamente, los estados de entidad surgen de la combinación de los atributos descriptivos para un rango de valores conocidos.

De este modo, a partir de la definición de estados de entidad y escenarios, la siguiente actividad define el modelo de transición. En dicha actividad, se puede definir el modelo de transición entre escenarios, estados de entidad, o la combinación de ellos (en un todo de acuerdo con el uso de escenarios y estados de entidad). Luego, se definen las métricas responsables de cuantificar los atributos de la categoría de entidad y las propiedades de contexto (cuando estuvieren presente). En este punto es posible reutilizar definiciones de métricas disponibles en el repositorio de experiencias previas.

La implementación de la métrica consiste en ejecutar la medición de acuerdo con la definición de ésta y obtener así un valor numérico (o distribución de valores con su respectiva probabilidad). Es decir, la implementación refiere a ejecutar el método de medición usando un instrumento para cuantificar y recolectar medidas (expresadas en una unidad y escala dada), que luego son interpretadas por los indicadores. Paralelamente a la implementación del proceso de medición, se diseña la evaluación, es decir, de qué modo se desea interpretar las medidas obtenidas a través de las métricas. Para ello, se especifican los criterios de decisión para interpretar los valores, considerando los estados de entidad y escenarios cuando estuvieren definidos. Así, los datos interpretados a partir del empleo de los criterios de decisión son derivados a la tubería de visualización.

La tubería de visualización de Munzner se articula con 1) Objetivo, límites, y alcances del proyecto de medición, 2) Especificación de las métricas, 3) Medidas generadas a partir de la definición de la métrica, 4) Especificación del indicador para evaluar los valores, y 5) Las interpretaciones surgidas del indicador a partir de las medidas. Con esta información, se lleva adelante el mapeo visual, prototipos, y estrategias de visualización se generan iterativamente para satisfacer los requisitos de medición y evaluación indicados por el usuario. Una vez que han sido implementados (Por ejemplo, utilizando PowerBI, Tableau, QlikSense, entre otros), se obtiene suficiente retroalimentación para analizar y recomendar las acciones basadas en la experiencia y conocimiento previo.

Ahora bien, la primera sección introdujo la extensión de la ontología, mientras que la presente describió los pasos necesarios para definir un proyecto de medición a partir de sus conceptos. La siguiente sección introduce un formato de datos autocontenido, consistente, y libre de etiquetas capaz de intercambiar múltiples definiciones de proyectos de medición alineados con GOCAME-ESVI y ECINCAMI entre diferentes dispositivos.

3.3 Optimizando el Intercambio de Proyectos: BriefPD

El proyecto de medición basado en la ontología extendida (ECINCAMI) se define mediante GOCAME-ESVI. Ahora bien, una vez que la definición es consensuada, esta debe ser intercambiada entre los diferentes dispositivos y actores del esquema de recolección de datos. Se propone un nuevo modo de organización del proyecto de medición basado en la ontología ECINCAMI denominado *BriefPD* [117]. Este nuevo

modo de organización es sintáctica y semánticamente interoperable, no emplea etiquetas, y es autocontenido [118]. Desde la perspectiva sintáctica, los conceptos son expresados a través del nuevo formato de datos. Desde la perspectiva semántica, cada elemento intercambiado refiere consistentemente a la ontología ECINCAMI descrita mediante OWL. Por tal motivo, esta sección se desglosa en dos. La primera describe la organización de los datos del proyecto de medición basado en la ontología ECINCAMI. La segunda sección analiza los resultados de la simulación del nuevo formato de datos contrastados con XML (eXtensible Markup Language) y JSON (JavaScript Object Notation).

3.3.1 Organización de Datos BriefPD

La Figura 15 describe los conceptos principales involucrados en la definición de un proyecto de medición. Esta nueva organización permite lograr un contenido a intercambiar mediante mensajes que es:

- **Integrado**: Es posible definir múltiples definiciones de proyectos de medición simultáneamente a través de un único mensaje, mientras que se incluyen todos los conceptos necesarios para implementar la medición,
- **Autocontenido**: La totalidad de la definición del proyecto de medición es organizada y contenida dentro de un mismo mensaje,
- **Jerárquicamente organizada**: Todos los conceptos se organizan siguiendo un orden de navegación conveniente basado en la ontología ECINCAMI (Ver Figura 12). Por ejemplo, a partir de una categoría de entidad bajo monitoreo es posible conocer los atributos característicos a cuantificar.
- **Consistente**: Cada concepto involucrado en la definición del proyecto tiene su correlato en la ontología ECINCAMI.

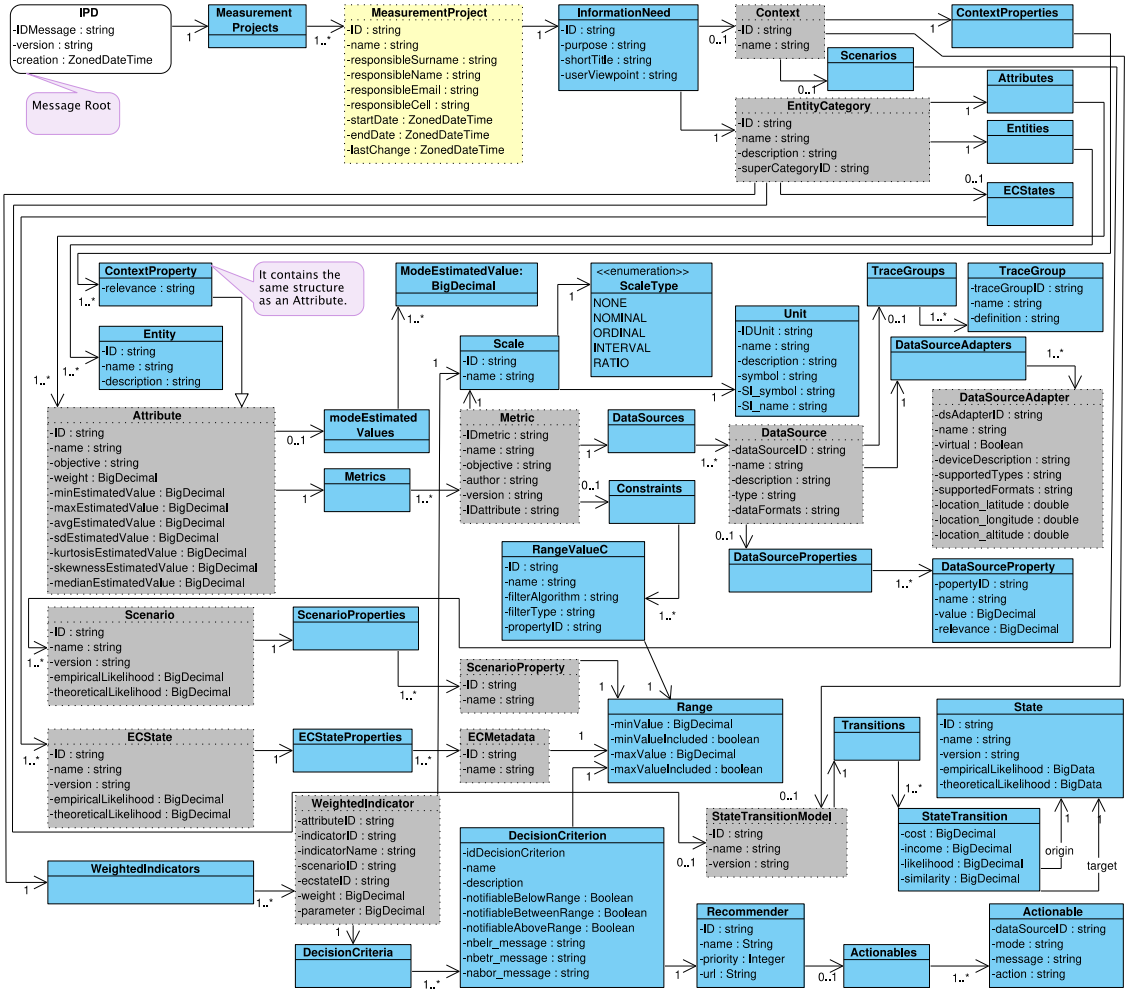

Figura 15 Conceptos Principales de la Definición de Proyecto Integrada. Una Descripción Autocontenida Basada en una Ontología de Medición

La nueva organización jerárquica puede informar un conjunto de proyectos de medición en forma simultánea. La raíz del mensaje se indica con un fondo blanco (Ver etiqueta *IPD* en Figura 15) y tiene asociada una etiqueta *MeasurementProjects* (Rectángulo con fondo amarillo y línea punteada) que permite agrupar un conjunto de proyectos de medición. Cada elemento dentro de la etiqueta *MeasurementProject* describe uno o más conceptos requeridos para el proyecto de medición (sintetizado en la Figura 12). El punto de partida para el proyecto es la descripción es la necesidad de información (elemento *InformationNeed*). A partir de allí, se describe la categoría de entidad a monitorear (elemento *EntityCategory*) y el contexto cuando estuviere presente (etiqueta *Context*). Bajo la categoría de entidad se definen los atributos (junto con el modo de cuantificarlos a través de métricas y fuentes de datos asociadas), entidades, estados de entidad, modelos de transición de estados, e indicadores ponderados (junto con los recomendadores asociados). Por otro lado, bajo el elemento del contexto se definen las propiedades contextuales (también indicando el modo de cuantificarlas mediante métricas y fuentes de datos asociadas), escenarios, y modelos de transición. Debido a que la propiedad de contexto es un atributo (Ver la

especialización en Figura 12), la interpretación de datos es realizada mediante los indicadores a través de los atributos que se cuantifican (por ejemplo, humedad ambiental).

El orden de navegación basado en la ontología ECINCAMI (Ver Figura 12) describe el modo en que los conceptos y dependencias son organizados. Esta organización jerárquica e integrada permite generar un mensaje conteniendo múltiples proyectos de medición concurrentemente.

El mensaje se genera de acuerdo con la organización jerárquica de conceptos introducida en la Figura 15. El punto de partida lo representa el elemento *IPD* representando la raíz del mensaje. Entonces, cada elemento que lo compone es alcanzado mientras desciende recursivamente a lo largo de la jerarquía. Cada atributo de clase (sea o no compuesto) es leído o escrito en un orden estricto definido por la clase. Cada clase incorpora un identificador que asocia el elemento de texto con su significado.

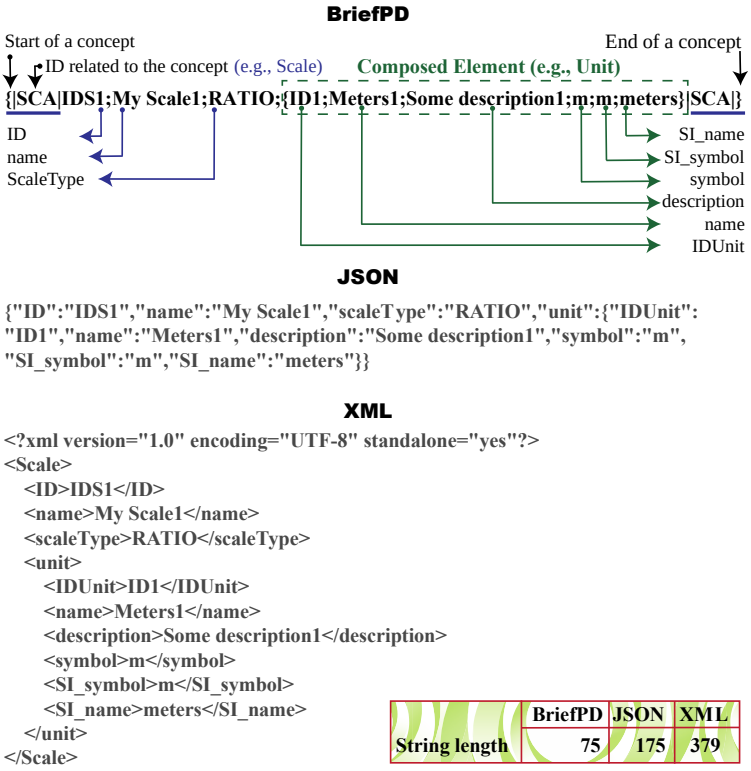

Figura 16 Un Fragmento Comentado del Mensaje Asociado con la Definición del Proyecto de Medición Integrada

La Tabla 14 describe los identificadores asociados con cada uno de los conceptos de la ontología ECINCAMI. De este modo, cuando se necesita escribir una instancia de la clase *Scale*, "{|SCA|" indica el inicio del concepto.

Tabla 14 Identificadores para los Conceptos descritos en la ontología ECINCAMI.

Class	ID	Class	ID
Actionable	ALE	InformationNeed	INF
Actionables	ALES	IPD	IPD
Attribute	ATT	MeasurementProject	MPR
Attributes	ATS	MeasurementProjects	MPS
Constraint	CO	Metric	ME
Constraints	COS	Metrics	MES
Context	CTX	Range	RGE
ContextProperties	CPS	Recommender	REC
ContextProperty	CP	Scale	SCA
DataSource	DS	Scenario	SCE
DataSources	DSS	ScenarioProperties	SPS
DecisionCriteria	DCA	ScenarioProperty	SP
DecisionCriterion	DC	Scenarios	SNS
ECMetadata	EC	StateTransition	ST
ECState	EST	StateTransitionModel	STM
ECStateProperties	ECS	Transitions	TRA
ECStates	STS	Values	VUS
Entities	ETS	WeightedIndicator	WIN
Entity	ENT	WeightedIndicators	WIS
EntityCategory	ECT		

El primer elemento luego del identificador de concepto corresponde al primer atributo de la clase, es decir, ID para Scale (Ver Figura 16). Luego, se escriben los atributos *name* y *scaleType* separados por ";". Como se puede observar en la Figura 16, *scaleType* es una enumeración y *Unit* corresponde con una instancia compuesta. Por ese motivo, luego del punto y como aparece una nueva llave izquierda "{" comenzando el nuevo concepto, pero sin indicar un ID. Esto se debe a que la instancia *Unit* no contiene atributos compuestos. Así, los atributos de *Unit* se escriben en orden separados por punto y coma. El punto y coma siempre se encuentra presente entre los atributos, aunque no al inicio o fin. De este modo, no se utilizan etiquetas para incorporar el valor de cada atributo (las únicas etiquetas corresponden con los identificadores de la clase). Este aspecto representa una ventaja en el tamaño del mensaje debido al ahorro que produce, como puede apreciarse en la Figura 16 contrastando el mismo mensaje entre los formatos BriefPD, JSON, y XML. Es decir, mientras que BriefPD requiere 75 caracteres para representar la relación entre escala, tipo de escala y unidad, XML y JSON requieren 379 y 175 caracteres respectivamente. Este ahorro se soporta en evitar usar etiquetas y aprovechando la jerarquía de conceptos.

BriefPD puede ser convertido hacia y desde representaciones JSON y XML. El contenido puede leerse sencillamente siguiendo las reglas de generación indicadas soportando la interoperabilidad sintáctica. A su vez, cada concepto es emparejado e

interpretado de acuerdo con una jerarquía soportada por la ontología ECINCAMI, soportando la interoperabilidad semántica.

Debido a que el mensaje generado es una jerarquía natural de conceptos guiado por una ontología, es posible seguir la idea de árbol de Merkle [119]–[122] para verificación de integridad. Cada nivel del árbol (o jerarquía) puede computar una huella recursivamente utilizando los datos del nivel. De este modo, la raíz de árbol o jerarquía obtiene una única huella calculada a partir del contenido del mensaje. Ello permite comparar dos mensajes o proyectos (de acuerdo con el nivel de concepto a comparar) para determinar si son iguales a través del contraste de sus respectivas huellas. Adicionalmente, dado dos definiciones del proyecto de medición organizadas de acuerdo con el orden de navegación introducido, es posible compararlas parcialmente para encontrar diferencias mediante las huellas por nivel. A partir de ello, se puede actualizar parcial e incrementalmente la definición del proyecto reemplazando parte del árbol o jerarquía.

3.3.2 Análisis de BriefPD basados en una Simulación Discreta

Se implementó BriefPD como prueba de conceptos basado en la ontología ECINCAMI y modelo de navegación derivado introducido anteriormente. A tal efecto, se empleó Java 8 e incorporó en la librería de código abierto *cincamipd*. Dicha librería es públicamente accesible a través de GitHub (https://github.com/mjdivan/cincamipd) y se liberó bajo los términos del acuerdo de licencia Apache 2.0. La librería escribe, lee, e intercambia definiciones de proyecto de medición en BriefPD como así también en JSON y XML. A su vez, es posible traducir una misma representación de proyectos de medición entre los mencionados formatos indistintamente. Se emplearon las librerías JAXB y GSON 2.8.6 para producir las representaciones en XML y JSON respectivamente. El nuevo formato puede comunicar el conjunto de proyectos de medición sin utilizar etiquetas y en forma autocontenida. Adicionalmente, se mantiene seguimiento de los conceptos a través del orden de generación derivado de la ontología.

Se desarrollaron tres simulaciones discretas. Los datos para las simulaciones se generaron utilizando la definición de un conjunto testigo de proyectos de medición basado en el caso de pacientes ambulatorios descrito en la

Figura 17.

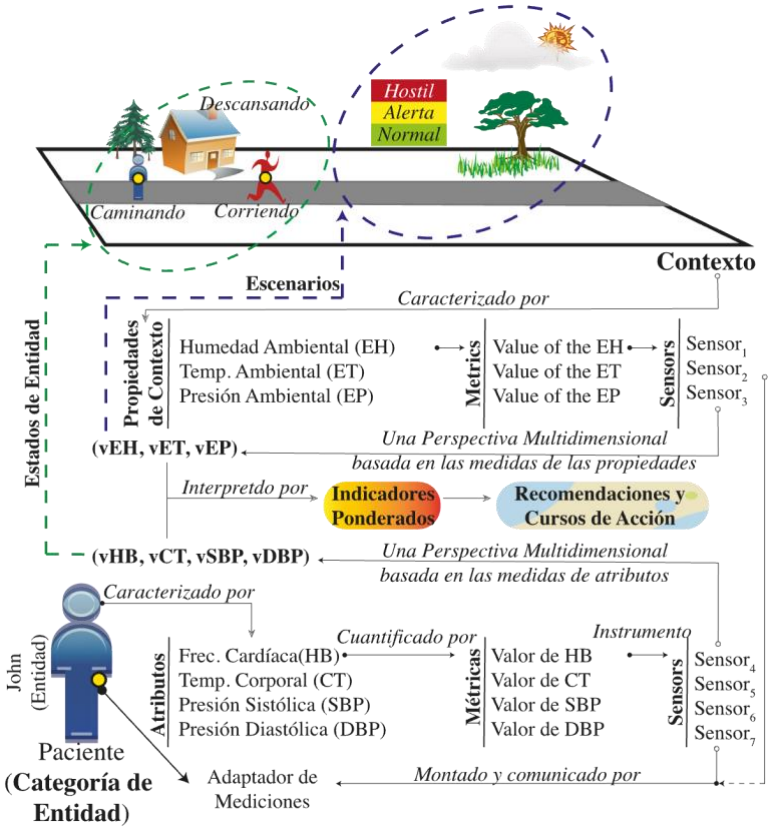

Figura 17 Ejemplo Conceptual basado en ECINCAMI para la Simulación

Sintéticamente, la figura introduce un paciente ambulatorio como categoría de entidad para quien se monitorean los atributos *frecuencia cardiaca*, *temperatura corporal*, *presión sistólica*, y *presión diastólica*. La entidad bajo análisis se denomina *John*. Cada atributo se cuantifica mediante su respectiva métrica la que se asocia a su vez con un sensor particular. La combinación de valores derivados de las métricas asociadas con los atributos, definen tres estados de entidad para identificar cuando la categoría de entidad se encuentra *corriendo*, *descansando*, o *caminando*. Se incorpora un contexto del cual se desea conocer su humedad ambiental, temperatura ambiental, y presión ambiental (es decir, sus propiedades de contexto). Los mismos se asocian con sus respectivas métricas y sensores que permiten obtener la medida. La combinación de valores para las propiedades contextuales define tres escenarios de análisis: *hostil*, *alerta*, y *normal*. Los indicadores ponderados definen sus criterios de decisión basados en los estados de entidad y escenarios, para luego asociarse con recomendaciones y-o cursos de acción. Así, el presente caso permite generar las definiciones de proyecto de medición para comparar BriefPD, JSON, y XML.

Utilizando un contenido estable (sin cambios), la primera simulación tuvo por objetivo verificar 1) La correcta traducción y mapeo de cada concepto a lo largo de la jerarquía ECINCAMI para un conjunto de proyectos en un mismo mensaje, 2) La sobrecarga incorporada por las operaciones de compresión y descompresión, 3) La

posibilidad de verificación parcial, y 4) La estimación de patrones de referencia (tiempo y tamaño) para BriefPD, XML, y JSON. El modelo de objetos se genera automáticamente en la simulación a partir del modelo de proyecto de referencia introducido en la Figura 17. Durante 10 minutos y manteniendo constante el número de proyectos por mensaje generado, se midió las operaciones de a) Generación del mensaje, b) Compresión del mensaje, c) Descompresión del mensaje, d) Regeneración del mensaje (desde el contenido representado como cadena de caracteres al modelo de objetos derivado de ECINCAMI), y e) Verificación parcial (es decir, buscar diferencias parciales en la jerarquía de conceptos).

Sin embargo, la segunda simulación tuvo por finalidad verificar las mismas operaciones que la primera simulación cuando el contenido del mensaje (es decir, los proyectos de medición a informar) se modifica dinámicamente. Se focaliza aquí sobre las variaciones potenciales de los patrones de referencia en cada formato de datos (es decir, BriefPD, JSON, y XML). Así, el número de proyectos por mensaje se varió durante la ejecución de 10 a 100 para medir el impacto sobre los tiempos y tamaños de operación según los formatos de datos.

La tercera simulación se focalizó en verificar la operación de verificación parcial para diferentes niveles de la jerarquía de conceptos de ECINCAMI. De este modo, una vez generada una definición de proyecto a comunicar, la estructura es intencionalmente modificada en diferentes niveles de la jerarquía para estimar las variaciones temporales por operación durante 2 minutos por cada concepto.

Tabla 15 Simulación 1. Patrones de Referencia de Tiempo y Tamaño

Concepto	Formato	Media	Mediana	Desvío Estándar
Tamaño del mensaje (KB)	XML	1417,17	1417,17	0,00
	JSON	371,69	371,69	0,00
	BriefPD	181,57	181,57	0,00
Tamaño Comprimido (KB)	XML	44,20	44,20	0,00
	JSON	17,21	17,21	0,00
	BriefPD	3,51	3,51	0,00
Tiempo de Generación (ms)	XML	76,82	74,93	12,72
	JSON	3,38	3,07	3,98
	BriefPD	7,51	6,95	2,48
Tiempo de Compresión (ms)	XML	567,28	567,05	12,46
	JSON	149,51	149,36	3,95
	BriefPD	72,54	72,26	2,35
Tiempo Total de Generación (ms)	**XML**	**644,09**	**642,36**	**19,78**
	JSON	**152,89**	**152,48**	**6,19**
	BriefPD	**80,05**	**79,54**	**3,88**

Tiempo de Descompresión (ms)	XML	3,62	3,53	0,40
	JSON	1,00	0,97	0,22
	BriefPD	0,48	0,44	0,35
Tiempo de Regeneración (ms)	XML	75,56	74,05	11,47
	JSON	3,65	3,43	1,54
	BriefPD	5,32	4,68	3,01
Tiempo Total de Recarga (ms)	**XML**	**79,18**	**77,69**	**11,62**
	JSON	**4,65**	**4,41**	**1,57**
	BriefPD	**5,80**	**5,14**	**3,27**

La Tabla 15 describe los patrones de referencia por operación, manteniendo el contenido del mensaje a intercambiar estable por 10 minutos. Desde la perspectiva del tamaño del mensaje, BriefPD consume 181,57 KB contra 371,69 KB y 1417,17 KB de JSON y XML. BriefPD mejora el tamaño consumido contra JSON 2,04 veces (Esto representa un 48,84% del tamaño de JSON) y contra XML 7,8 veces (lo que implica un 12,81% del tamaño de XML). Una situación similar ocurre con el contenido comprimido. BriefPD consume 3,51 KB versus 17,21 KB y 44,20 KB de JSON y XML. BriefPD optimiza el tamaño comprimido contra JSON 4,9 veces (Esto es un 20,39% del tamaño de JSON) y contra XML 12,59 veces (lo cual implica un 7,94% del tamaño de XML). Así, BriefPD intercambia un conjunto de definiciones en un mensaje simple y libre de etiquetas, consistentemente basado en la ontología ECINCAMI. Se optimiza el tamaño del contenido contra el mismo expresado en JSON y XML. Como resultado de esta simulación, se concluye que sería posible expresar definiciones basadas en la mencionada ontología bajo BriefPD, JSON, o XML a demanda. Como patrón de referencia, 10 proyectos de medición testigos expresados en BriefPD consumirían 181,57 KB (o 3,51 KB comprimido) sin apreciarse variaciones significativas en su tamaño.

Desde la perspectiva del tiempo de generación total para un mensaje con 10 proyectos de medición, BriefPD requiere 80,05 ms versus 152,89 ms y 644,09 ms de JSON and XML. BriefPD hace mejor uso del tiempo en contraste a JSON 1,9 veces (lo que representa un 52,35% del tiempo de JSON) y versus XML 8,04 veces (lo que implica un 12,42% del tiempo de XML). Sin embargo, JSON obtiene mejores resultados cuando el mensaje se recarga desde la representación plana. Es decir, BriefPD requiere 5,8 ms versus 4,65 ms y 79,18 ms de JSON and XML. BriefPD necesita 24.73% más de tiempo que JSON. Pero, este es 13,65 veces mejor que XML (lo que implica 7.32% del tiempo de XML). Esto esboza un desafío a abordar en trabajos futuros. Por un lado, y como patrón de referencia, el tiempo total para generar un mensaje BriefPD con 10 proyectos consumiría 80,05 ms con una desviación estándar de 3,88 ms. Por otro lado, el tiempo total para regenerar el modelo conceptual derivado de ECINCAMI desde una representación de texto plana consumiría 5,80 ms con una desviación estándar de 3,17 ms.

La Tabla 16 describe el resultado de la segunda simulación para el tiempo de generación cuando su contenido varía de 11 a 101 definiciones de proyecto por mensaje. Las primeras tres columnas indican el tiempo absoluto consumido por BriefPD, JSON, y XML respectivamente. Los mejores resultados de BriefPD sobre JSON y XML observados en la primera simulación son mantenidos a lo largo de toda la segunda simulación. BriefPD requiere 790,82 ms cuando el mensaje contiene 101 definiciones de proyectos de medición en contraste con 1503,02 y 5938,53 de JSON y XML. BriefPD hace un mejor uso del tiempo en comparación a JSON 1,9 veces (52,61% del tiempo de JSON) y XML 7,5 veces (13,31% del tiempo de XML). Las últimas tres columnas describen la evolución del tiempo promedio por proyecto requerido para generar la representación correspondiente. De este modo, BriefPD requiere de 7,37 ms por proyecto con una desviación estándar de 2,55 ms contra 13,9 ms (Desviación estándar de 4,91) y 54,96 ms (Desviación estándar de 16,22) de JSON y XML.

Tabla 16 Simulación 2. Evolución del Tiempo de Generación (ms)

#	BriefPD	JSON	XML	$\overline{x}_{BriefPD}$	$\overline{x}_{JSON}$	$\overline{x}_{XML}$
11	155,43	299,14	1026,83	14,13	27,19	93,35
21	182,60	284,71	1483,82	8,70	13,56	70,66
31	204,94	379,47	1475,02	6,61	12,24	47,58
41	263,44	467,81	1875,64	6,43	11,41	45,75
51	312,07	590,68	2281,98	6,12	11,58	44,74
61	365,13	674,77	2720,53	5,99	11,06	44,60
71	425,71	789,51	3076,80	6,00	11,12	43,34
81	484,25	894,86	3494,34	5,98	11,05	43,14
91	541,09	1355,06	5243,97	5,95	14,89	57,63
101	790,82	1503,02	5938,53	7,83	14,88	58,80

La Tabla 17 sintetiza los resultados de la simulación 2 para el tiempo de recarga del mensaje cuando el número de proyectos varía entre 11 y 101.

Tabla 17 Simulación 2. Evolución del Tiempo de Recarga (ms)

#	BriefPD	JSON	XML	$\overline{x}_{BriefPD}$	$\overline{x}_{JSON}$	$\overline{x}_{XML}$
11	73,56	37,21	244,04	6,69	3,38	22,19
21	44,84	30,97	231,34	2,14	1,47	11,02
31	37,83	22,69	192,02	1,22	0,73	6,19
41	45,03	21,28	172,95	1,10	0,52	4,22
51	44,82	23,28	174,81	0,88	0,46	3,43
61	50,32	27,93	184,02	0,82	0,46	3,02
71	55,59	31,69	209,01	0,78	0,45	2,94
81	79,02	34,58	223,86	0,98	0,43	2,76
91	73,12	39,02	235,27	0,80	0,43	2,59
101	86,70	62,08	271,67	0,86	0,61	2,69

Las primeras tres columnas describen el tiempo absoluto para regenerar el modelo de objetos a partir del mensaje representado como texto plano. Se confirma es más ágil que BriefPD para regenerar el modelo de objetos a lo largo de la simulación, relegando a XML como tercera alternativa. BriefPD requiere 86,70 ms para regenerar el modelo de 101 proyectos en memoria contra 62,08 ms y 271,61 ms de JSON y XML. De este modo, BriefPD consume un 39,65% de tiempo adicional contra JSON pero es 3,13 veces menor que XML (31,91% del tiempo de XML). Esto da la oportunidad para mejorar el proceso de regeneración en BriefPD como trabajo futuro. Las últimas tres columnas indican la variación de la media aritmética a lo largo de la simulación y composición de proyectos por mensaje. Así, BriefPD requiere de 1,63 ms con desviación estándar de 1,82 ms para recargar un proyecto, mientras que JSON y XML necesitan 0,89 ms (Desv.Est. 0,93) y 6,10 ms (Desv. Est. 6,22). Debido al recolector de basura de Java, se identifican valores atípicos (o outliers) en las tasas, lo cual puede constatarse contrastando la media aritmética respecto de la media recortada al 95%.

Tabla 18 Simulación 3. Resultados de la Verificación Parcial y Actualización del Proyecto de Medición (ms)

#	Verificación	Recarga	Actualización Parcial	Tiempo de Respuesta
DataSourceAdapter	0,0039	0,0003	0,0001	0,0043
ECMetadata	0,0057	0,0004	0,0001	0,0062
ScenarioProperty	0,0061	0,0005	0,0001	0,0067
DataSource	0,0069	0,0007	0,0001	0,0077
Scenario	0,0149	0,0018	0,0001	0,0168
Metric	0,0188	0,0023	0,0001	0,0213
Attribute	0,0226	0,0036	0,0001	0,0263
ECState	0,0241	0,0033	0,0002	0,0275
StateTransitionModel	0,0415	0,0035	0,0002	0,0452
WeightedIndicator	0,1057	0,0120	0,0003	0,1180
Context	0,2105	0,0371	0,0022	0,2498
EntityCategory	1,1593	0,1779	0,0125	1,3497
MeasurementProject	1,4610	0,3578	0,0007	1,8195

La Tabla 18 describe el tiempo involucrado en verificar junto con el tiempo total de respuesta para la verificación parcial del mensaje en diferentes niveles conceptuales de navegación para ECINCAMI (Ver Figura 15). La misma se encuentra ordenada de acuerdo con el tiempo de respuesta en forma ascendente. De este modo, utilizando la organización jerárquica basada en la ontología sustentado en la idea del árbol de Mekle, se miden las operaciones de verificación, recarga (o regeneración), reemplazo (o actualización parcial), junto con el tiempo de respuesta. Así, es posible indicar que la verificación corresponde con la operación más costosa independientemente del nivel conceptual, seguido por la regeneración y actualización parcial en dicho orden. El tiempo

de respuesta varía entre 0,0043 ms y 1,8195 ms de acuerdo con el nivel jerárquico de cada concepto. Tales tiempos son pertinentes al entorno de Internet de las Cosas (IoC o IoT por su acrónimo en inglés para Internet of Things). Esta funcionalidad facilita la oportunidad de incorporar actualizaciones parciales (consumiendo hasta 1,8195 ms) en dispositivos de bajo recursos que requieran implementar la actualización.

3.4 Conclusión Generales del Capítulo

El presente capítulo introdujo ECINCAMI como extensión ontológica para la gestión de nuevos conceptos, términos y relaciones en un proyecto de medición. ECINCAMI se encuentra publicada y accesible a través de la web, incorporando aspectos de especial interés al proceso de medición:

- La conceptualización de las fuentes de datos y grupos de seguimiento para facilitar la reutilización de medidas desde diferentes perspectivas,
- La asociación de las fuentes de datos con restricciones y propiedades que permiten una mejor selección al momento del emparejamiento con las métricas,
- La gestión de medidas deterministas y no deterministas para gestionar situaciones en los cuales un único valor numérico sería muy costoso de obtener o no estuviere disponible para un cierto nivel de exactitud y precisión,
- La incorporación de los datos complementarios a la medida que permiten brindar elementos adicionales de análisis cuando sean requeridos,
- La gestión de estados de entidad para modelar situaciones observables de interés respecto a la categoría de entidad que se monitorea,
- La gestión de escenarios para modelar configuraciones observables del contexto de especial interés respecto a cómo podrían afectar (directamente o no) a la categoría de entidad bajo análisis,
- La posibilidad de definir indicadores cuya interpretación puede configurarse en términos de la combinación de estados de entidad y escenarios,
- La posibilidad de asociar indicadores con recomendadores para guiar la selección de cursos de acción respecto a una interpretación particular,

La estrategia GOCAME-ESVI permite articular el proceso de definición de proyectos de medición respecto a los nuevos conceptos ontológicos introducidos. Ello permite guiar al usuario a través de una serie de pasos guiados por la necesidad de información, objetivo y alcance para obtener la definición del proyecto de medición.

Se introdujo un nuevo formato para el intercambio de proyectos de medición denominado BriefPD. Este formato es jerárquicamente organizado basado en el modelo de navegación de conceptos derivado de ECINCAMI, autocontenido, libre de etiquetas, verificable y actualizable parcialmente siguiendo la estrategia derivada del árbol de Merkle. Desde la perspectiva sintáctica, reglas simples de generación permiten

intercambiar contenido sin etiquetas basado en el modelo de navegación derivado de ECINCAMI optimizando el tamaño. Desde la perspectiva semántica, cada concepto intercambiado en los proyectos de medición se sustenta en ECINCAMI. Dicha ontología es definida mediante OWL y públicamente accesible mediante repositorios. Adicionalmente, la librería cincamipd disponible a través de GitHub y licencia Apache 2.0, provee una implementación de referencia para BriefPD, como así también de traducir el mismo hacia y desde JSON y XML.

Las simulaciones discretas desarrolladas proveen un patrón de referencia en términos de tiempos requeridos de generación (Ver Tabla 19), verificación, recarga, actualización, e intercambio para entornos que requieran procesamiento de datos en tiempo real. Limitado a las simulaciones, es posible indicar que BriefPD fue la mejor opción en términos de tamaño y tiempo de generación (es decir, representó el 20,39% del tamaño de JSON y se generó 1,9 veces más rápido que JSON). Se obtuvieron resultados interesantes en el proceso de recarga que serían una oportunidad de mejora a futuro. De este modo, como patrones de referencia para su aplicabilidad, puede sintetizarse la siguiente información desde las simulaciones discretas:

Tabla 19 BriefPD. Referencias de Tiempos y Tamaños de Procesamiento

Concepto	Referencia
BriefPD – T. de generación del mensaje con 10 proyectos	80,05 ms
BriefPD – T. de recarga para un mensaje con 10 proyectos	5,8 ms
BriefPD – Tamaño de mensaje sin compresión para 10 proyectos	181,57 KB
BriefPD – Tamaño de mensaje comprimido para 10 proyectos	3,51 KB
BriefPD – T. de generación por proyecto. Media recortada (95%)	6,27 ms
BriefPD – T. de recarga por proyecto. Media recortada (95%)	0,93 ms
BriefPD – T. de respuesta en verificación/actualización parcial	[0,0045; 1,8195] ms

Como síntesis complementaria a los patrones de referencia, se proveen los siguientes recursos:

- ECINCAMI está descrita mediante OWL (Ontology Web Language) y se encuentra disponible bajo los términos de la licencia *Creative Commons 4.0* en *Figshare* y en el *Repositorio y Registro de Ontologías*.
- La librería *cincamipd* provee una implementación de referencia de *BriefPFD,* es públicamente accesible a través de GitHub (https://github.com/mjdivan/cincamipd) bajo los términos de la licencia Apache 2.0
- Una cápsula del servicio Code Ocean se hizo públicamente disponible y accesible para garantizar el acceso y reproducibilidad del experimento. Se ha dispuesto del código y software necesario para su ejecución en la nube (https://codeocean.com/capsule/4500824/tree/v1).

El siguiente capítulo aborda el cómputo de medidas de similitud estructural y comportamental basado en los conceptos ontológicos de ECINCAMI. Este aspecto toma particular importancia al momento de determinar cuan parecidos son dos proyectos entre sí a los efectos de reutilizar experiencias o conocimientos previos entre ellos.

Capítulo 4	Medidas de Similitud para Proyectos de Medición y Evaluación

Capítulo 4. Medidas de Similitud para Proyectos de Medición y Evaluación

Introducción

El capítulo anterior ha introducido ECINCAMI, una ontología de medición y evaluación extendida que incorporaba nuevos conceptos y relaciones tales como medidas probabilistas y deterministas, estados de entidad, escenarios como estados de los contextos, fuentes de datos, grupos de seguimientos, datos complementarios, indicadores ponderados basados en escenarios y estados de entidad, entre otros aspectos.

A partir de la estrategia extendida GOCAME-ESVI, es posible emplear ECINCAMI para describir y formalizar un proyecto de medición y evaluación considerando aspectos y guías de visualización. De este modo, se alinea desde la necesidad de información que guía la definición del proyecto, hasta el cómo visualizar las diferentes métricas e indicadores en forma consistente.

Se introdujo un nuevo formato denominado BriefPD. Éste permite intercambiar las definiciones de proyectos de medición y evaluación en forma consistente. El formato es interoperable sintácticamente dado que existen un conjunto de pasos simples para su generación y lectura a partir del modelo de navegación derivado de la ontología ECINCAMI. A su vez, es semánticamente interoperable debido a que todo concepto intercambiado se sustenta en un elemento disponible en la mencionada ontología. El formato es jerárquicamente organizado a partir del modelo de navegación jerárquico derivado de ECINCAMI, lo que permite verificación parcial del contenido a través de un abordaje similar al de los árboles de Merkle. Finalmente, el formato es autocontenido y puede representar un conjunto de proyectos de medición dentro de un mismo mensaje o contenido.

De este modo, es posible el intercambio de definiciones de proyectos de medición entre diferentes componentes del sistema de recolección, lo que permite que cada uno conozca su rol en la recolección, como así también la relación entre sensor, el valor numérico, medida, y el atributo de una entidad dada.

En cada proyecto, los criterios de decisión utilizan conocimiento de expertos y experiencias previas para interpretar un valor numérico (es decir, la medida) a través de los indicadores, y obtener así, un clasificador adecuado para proveer recomendaciones. Estos criterios se relacionan con los sistemas de recomendación, quienes proveen un conjunto de cursos de acción mediante clasificadores [123]. Cuando un proyecto de medición es nuevo o existe poca (o ninguna) información respecto de ciertas situaciones, no existe conocimiento previo para obtener un clasificador adecuado en el proyecto actual, lo que podría derivar en la ausencia de recomendación.

Desde la perspectiva de la teoría de la información, dos conceptos son similares cuando ellos comparten un número de propiedades comunes. Así, la similitud podría

representarse como una magnitud entre 0 y 1. Cuando un valor tiende a estar cerca de uno, se dice que ambos conceptos tienden a ser similares. Sin embargo, cuando el valor es cercano a cero, se dice que los conceptos tienden a ser diferentes o no similares. Pero la similitud semántica considera el significado del contenido para aproximar dicha similitud y no solo la mera intersección de propiedades. Existen diferentes métodos para estimar y calcular similitudes como se describe en [124].

Dado que los proyectos de medición se intercambian mediante BriefPD y cada elemento se sustenta en una base conceptual común de la ontología ECINCAMI, el significado subyacente para los diferentes conceptos de proyectos ha sido definido. Así, conociendo las definiciones de diferentes proyectos de medición (plausibles de intercambio mediante BriefPD), el eje del capítulo reside en cómo estimar la similitud entre proyectos de medición junto con una estimación del coste de hacerlo. A partir de dicha estimación de similitud, es posible ordenar los proyectos similares en forma descendente para localizar clasificadores alternativos cuando un proyecto dado no cuenta con suficiente información para hacerlo.

Los sistemas de toma de decisiones en tiempo real requieren tomar decisiones tan pronto un nuevo dato arriba para su procesamiento [125]. Esto representa un desafío debido a que el procesamiento de datos en sí mismo debe darse con los recursos disponibles en ese instante de tiempo. Este es un aspecto importante debido a que los proyectos de medición tienden a ser automatizados y alimentar diferentes sistemas en tiempo real [126]. De este modo, un proyecto de medición podría estar recibiendo datos desde los sensores (es decir, la fuente de datos), pero el clasificador asociado (responsable de localizar cursos de acción basado en los criterios de decisión y datos de los indicadores) podría estar inicializándose o no disponer aún de suficientes datos. Esta situación se conoce como el desafío de inicio en frío (en inglés, "*Cold Start*") en los sistemas de recomendación de filtrado colaborativo [127].

En términos de ECINCAMI, una entidad perteneciente a una categoría de entidad podría experimentar diferentes estados de entidad en interacción con diferentes escenarios. Ello implica que un valor numérico puede ser interpretado de modo diferente de acuerdo con el estado de entidad y escenario actual. Tanto la determinación del escenario actual como del estado de entidad se realiza a partir de la lectura de las medidas que los define [128]. De este modo, un seguimiento en línea de los sensores vinculados con las métricas (asociadas con los atributos o propiedades de contexto), permite conocer las medidas respectivas y determinar en línea el estado y escenario actual para la entidad o contexto respectivamente. Así, identificado el estado y escenario para un indicador dado, las medidas se interpretan y sería posible obtener un criterio de decisión con su respectivo clasificador incremental que permita buscar recomendaciones de acuerdo con los valores interpretados.

Ahora bien, cuando la situación es nueva, puede que no existan cursos de acción o clasificadores capaces de proveer recomendaciones. No obstante, podría existir un

conjunto de datos históricos de otros proyectos de medición útiles para ayudar a direccionar la falta de conocimiento específico. Sin embargo, tal utilidad estaría sujeta a que las entidades monitoreadas sean similares en términos del objetivo de medición. Entonces, cuando un proyecto no tiene cursos de acción (o recomendaciones) para un criterio de decisión en particular de un indicador, un ordenamiento basado en similitud podría guiar la búsqueda de clasificadores pertinentes entre proyectos de medición heterogéneos para mitigar el riesgo de inicio en frio (o ausencia de información).

Este capítulo introduce las medidas de similitud basado en definiciones de proyectos organizadas en BriefPD basadas en ECINCAMI. Se aborda el análisis desde la perspectiva estructural y comportamental. La perspectiva estructural analiza la similitud para la definición intercambiada del proyecto que especifica entidades, métricas, y demás conceptos. La perspectiva comportamental, analiza la similitud asociada con los valores recibidos para las medidas e indicadores a partir de la definición desde los sensores. De este modo, dos proyectos podrían ser estructuralmente idénticos, aunque su comportamiento a partir de los sensores ser completamente diferente. Por ejemplo, se podría estar monitoreando el material particulado y sus efectos en la salud exactamente con la misma definición de proyecto de medición, aunque los datos procesados por la estación de monitoreo en la ciudad de Santa Rosa (La Pampa) serían completamente diferentes a los de la ciudad de Córdoba para el mismo intervalo [t; t+1]; lo cual es lógico debido a que se trata de dos regiones diferentes.

El capítulo se organiza en cinco secciones. La primera sección describe el abordaje de la similitud estructural. La segunda sección analiza la similitud comportamental. La tercera sección analiza conjuntamente las similitudes estructurales y comportamentales a través de un índice compuesto. La cuarta sección describe resultados de una simulación discreta que sirven como patrón de aplicabilidad. Finalmente, se esbozan conclusiones respecto del capítulo y sus contribuciones a este trabajo.

El capítulo se soporta en las siguientes publicaciones efectuadas a lo largo del proceso de investigación:

- Diván, M. Sánchez Reynoso, M. Méndez, M. Panebianco J. (2021) "**IoT-based Approaches for Monitoring the Particulate Matter and its Impact on Health**". e-ISSN: 2327-4662 – IEEE Internet of Things Journal. Institute of Electrical and Electronics Engineers Inc. 2021. Vol. 8, nro 15. pp. 11983 - 12003 http://dx.doi.org/10.1109/JIOT.2021.3068898

- Diván, M. Sánchez Reynoso, M. (2021) "**A Metadata and Z Score-based Load-Shedding Technique in IoT-based Data Collection Systems**". International Journal of Mathematical, Engineering and Management Sciences. ISSN: 2455-7749. e-ISSN: 2455-7749. Elsevier. Vol.6, nro 1. pp 363 – 382. https://ijmems.in/volumes/volume6/number1/23-IJMEMS-SBS19-34-6-1-363-382-2021.pdf

- Diván, M. Sánchez Reynoso, M. (2021) "**Strategies based on IoT for supporting the decision making in Agriculture: A Systematic Literature Mapping**". ISSN: 1755-0556 - e-ISSN: 1755-0564 - International Journal of Reasoning-based Intelligent Systems. Vol. 13, nro 3. pp155 – 171. https://dx.doi.org/10.1504/IJRIS.2021.117080

- Diván, M & Sánchez Reynoso, M (2020) "**A Real-Time Entity Monitoring based on States and Scenarios**" CLEI Electronic Journal. ISSN 0717-5000. Vol. 23 (1). Pp 2-1:2-25. https://doi.org/10.19153/cleiej.23.1.2

- Diván, M & Sánchez Reynoso, M (2020) "**Optimizing Data Transmission from IoT devices through Weighted Online Data Changing Detectors**" Advances in Data Science and Adaptive Analysis. ISSN 2424-922X. Vol 12 (2). 2041001 (pp.1:33). https://doi.org/10.1142/S2424922X20410016

- Sánchez Reynoso, M & Diván, M (2020) "**Assessment of semantic similarity in entities under monitoring: A systematic literature mapping**". Revista Facultad de Ingeniería Universidad de Antioquía. ISSN 0120-6230. Vol 99. Pp 21-31. https://doi.org/10.17533/udea.redin.20200476

- Diván, M & Sánchez Reynoso, M (2020) "**Relocating the Load-Shedding Strategy in the Data Stream Processing Architecture**" In IEEE 2020 Argencon. Resistencia, Chaco. 2 al 4 de diciembre. http://dx.doi.org/10.1109/ARGENCON49523.2020.9505446

- Sánchez Reynoso, M &Diván, (2020) "**Recomendación por similitud semántica en repositorios con grandes volúmenes de datos de medición**". V Jornadas de Intercambio y Difusión de los Resultados de Investigaciones de los Doctorandos de Ingeniería. UTN Córdoba, Córdoba. 6 y 7 octubre. ISBN: 978-950-42-0200-4. https://doi.org/10.33414/ajea.5.751.2020

- Sánchez Reynoso, M & Diván, M (2019) "**Improving the Real-Time Searching in the Organizational Memory**". Procedia Computer Science. Elsevier Ltd. Vol. 154, pp. 293-304. ISSN: 1877-0509. https://doi.org/10.1016/j.procs.2019.06.043

- Diván, M & Sánchez Reynoso (2019) "**A Load-Shedding Technique based on the Measurement Project Definition**". In V. Jain, S. Patnaik, F. Popentiu Vladicescu, and I.K. Sethi (Eds.). Proceedings of 5th International Conference on Intelligent Computing, Communication & Devices (ICCD 2018), Xi'an, China, November 22-

24 of 2018. In Advances in Intelligent Systems and Computing, Springer Nature Singapore. pp.1027-1033. ISSN 2194-5357. https://doi.org/10.1007/978-981-13-9406-5_122

- Sánchez Reynoso, M &Diván, (2019) "**A Systematic Literature Mapping on the Similar Semantically Entities in Measurement Projects**". 2019 International Conference on Virtual Reality and Visualization (ICVRV). Hong Kong, China. November 21-22 of 2019. https://doi.org/10.1109/ICVRV47840.2019.00033

- Diván, M & Sánchez Reynoso, M (2018) "**The Real-Time Measurement and Evaluation as System Reliability Driver**". Book Chapter in "System Reliability Management: Solutions and Technologies". Anand, A & Ram, M (Eds.). CRC Press, Taylor & Francis Group. Pp. 161-188. https://doi.org/10.1201/9781351117661-11

4.1 Similitud Estructural

En términos de ECINCAMI, un proyecto de medición define una categoría de entidad a monitorear alineado con el objetivo y necesidad de información. Esta categoría de entidad (ej., paciente ambulatorio) puede tener un conjunto de diferentes entidades asociadas (ej., Laura, Pedro, etc.). La categoría de entidad se caracteriza a través de un conjunto de atributos que son comunes a las entidades de este tipo. La definición estructural indica el conjunto de atributos monitoreados para una entidad dada (o el conjunto de propiedades contextuales para el contexto) con sus respectivos estados de entidad (o escenarios).

Sin embargo, dos entidades podrían tener la misma definición estructural (es decir, se monitorean mediante los mismos atributos), pero la actividad de cada una en diferentes escenarios podría derivar comportamientos diferentes. Es decir, la evolución de la frecuencia cardíaca para Laura y Pedro puede ser totalmente diferente de acuerdo con su actividad. De este modo, mientras la similitud estructural analiza los atributos y propiedades definidas para monitorear la categoría de entidad, la similitud comportamental analiza el comportamiento de la distribución de datos para cada entidad.

Sea e_i el conjunto finito y no vacío de atributos describiendo una categoría de entidad, $|e_i|$ representa el número de atributos que componen el conjunto. Análogamente, c_i es un conjunto finito y no vacío de propiedades contextuales (cp) que describen el contexto, mientras $|ci|$ representa el número de propiedades de contexto que los describen. Aquí, el índice de similitud tiene por objetivo calcular el nivel de coincidencia entre todos los atributos disponibles en ambos conjuntos finitos. La coincidencia implicaría que un atributo se encuentra presente en ambos conjuntos comparados expresado como vectores binarios (es decir, 1 indicaría presencia mientras

que 0 ausencia). Debido a que la frecuencia de los atributos es limitada a 0 y 1 (ausencia y presencia), la distancia coseno no es adecuada para este caso [124], [129]. Sin embargo, el coeficiente de Sorensen-Dice podría ser útil, aunque no satisface directamente la desigualdad triangular [130]. Por esta razón, se considera el índice Jaccard el cual expresa el nivel de coincidencia entre dos conjuntos discreto, finito, y no vacíos, satisfaciendo la desigualdad triangular [131]. Así, basado en el índice de Jaccard [132], la similitud estructural para dos entidades (es decir, e_1 and e_2) se indica en la Ecuación 1 como un cociente entre atributos comunes de entidades dividido por la unión de ambos conjuntos sin repetición.

Ecuación 1 Similitud Estructural de Entidades

$$sim_{str}^{ent}(e_1, e_2) = \frac{|e_1 \cap e_2|}{|e_1 \cup e_2|}$$

Cada conjunto contiene un número preciso y reducido de atributos escogido por expertos para monitorear una entidad (durante la definición del proyecto). La complejidad del monitoreo se equilibra con el número de atributos involucrados. Así, el índice Jaccard se torna en una simple pero interesante opción para calcular la similitud.

Un estado es entendido como una combinación bien establecida de atributos de entidad que podrían variar en un intervalo dado. Por ejemplo, el estado ´caminando´ en una entidad podría suponer un rango de variación para la frecuencia cardíaca, temperatura corporal, y la presión arterial de acuerdo con el ejemplo introducido en la Figura 17. Tal combinación de intervalos para diferentes atributos se agrupa bajo la idea de estado. Dado que diferentes combinaciones entre atributos e intervalos podrían definirse para una entidad, se puede especificar un conjunto de estados diferentes. La idea de transición representa la posibilidad de moverse desde un estado a otro (por ejemplo, transitar desde entidad caminando a corriendo, siguiendo los valores de los atributos monitoreados).

Sea $S(e_i)$ el conjunto que contiene el número de estados definidos para la entidad ´i´. Las transiciones entre estados para la misma entidad se indican como $T(e_i)$. Así, la similitud estructural para los estados de entidad de dos entidades (es decir, $S(e_1)$ y $S(e_2)$ para las entidades e_1 y e_2 respectivamente) se define en Ecuación 2, considerando un parámetro α que expresa la importancia relativa entre estados y sus transiciones. Es un valor limitado al intervalo cerrado [0; 1] que indica la proporción de importancia para los estados de entidad sobre las transiciones.

Ecuación 2 Similitud Estructural para los Estados de dos Entidades

$$sim_{str}^{st}(\alpha, e_1, e_2) = \alpha * \frac{|S(e_1) \cap S(e_2)|}{|S(e_1) \cup S(e_2)|} + (1-\alpha) * \frac{|T(e_1) \cap T(e_2)|}{|T(e_1) \cup T(e_2)|}$$

A partir de la Ecuación 1 y de la Ecuación 2, se define la distancia estructural interna como se indica en la Ecuación 3.

Ecuación 3 Distancia Estructural Interna

$$idist_{str}(\alpha, \beta, e_1, e_2) = 1 - \ [\beta * sim_{str}^{ent}(e_1, e_2) + (1 - \beta) * sim_{str}^{st}(\alpha, e_1, e_2)]$$

El parámetro β describe la importancia relativa de la entidad respecto de sus estados, asumiendo un valor en el intervalo [0; 1]. La anterior ecuación expresa una medida para saber cuán cerca dos entidades están entre ellas. De este modo, valores cercanos a cero representan entidades cercanas, mientras que valores cercanos a 1 representan entidades diferentes.

Desde la perspectiva estructural externa debe considerarse el rol del contexto y sus escenarios. De este modo, $S(c_i)$ indica un conjunto de escenarios para el contexto c_i, mientras $T(c_i)$ representa el conjunto de las transiciones para el contexto c_i. Siguiendo el razonamiento análogo para las entidades, la similitud estructural para dos contextos dados (es decir, c_1 y c_2) se define en la Ecuación 4.

Ecuación 4 Similitud Estructural de Contextos

$$sim_{str}^{ctx}(c_1, c_2) = \frac{|c_1 \cap c_2|}{|c_1 \cup c_2|}$$

Cada conjunto contiene un número preciso y reducido de propiedades contextuales especificado por expertos para monitorear un el contexto de una entidad (durante la definición del proyecto). La complejidad del monitoreo se equilibra con las propiedades contextuales involucrados. Un escenario se enciente como una combinación bien establecida de propiedades del contexto podrían variar conjuntamente en un intervalo dado. En este punto, las transiciones modelan los cambios de escenarios para un contexto dado a lo largo del tiempo.

Ecuación 5 Similitud Estructural de Escenarios

$$sim_{str}^{sc}(\gamma, c_1, c_2) = \gamma * \frac{|\ S(c_1) \cap\ S(c_2)|}{|S(c_1)\ \cup\ S(c_2)|} + \ (1 - \gamma) * \frac{|T(c_1)\ \cap\ T(c_2)|}{|T(c_1)\ \cup\ T(c_2)|}$$

El parámetro γ describe la importancia relativa de los escenarios respecto de sus transiciones. Este asume un valor dentro del intervalo cerrado [0; 1].

Ecuación 6 Distancia Estructural Externa

$$edist_{str}(\gamma, \delta, c_1, c_2) = 1 - \ [\delta * sim_{str}^{ctx}(c_1, c_2) + (1 - \delta) * sim_{str}^{sc}(\gamma, c_1, c_2)]$$

El parámetro δ indica la importancia relativa del contexto respecto sus escenarios. Este puede asumir un valor dentro del intervalo [0; 1] para representar la magnitud de su impacto, siendo un valor de 1 una incidencia completa del contexto en detrimento de sus escenarios, y 0 una incidencia completa de los escenarios en detrimento del contexto. Cualquier valor entre medio balancearía las influencias para evitar extremos.

De este modo, la Ecuación 6 permite expresar una distancia para saber cuán cerca estructuralmente se encuentran dos contextos entre sí. Un valor de cero indicaría dos contextos muy cercanos (o similares estructuralmente), mientras que un valor cercano a 1 indicaría fuertes diferencias entre ellos.

Hasta aquí, tanto la distancia estructural interna como externa refieren a aspectos que han sido definidos por expertos en el proyecto de medición basado en la ontología ECINCAMI y la entidad a monitorear. Sin embargo, esta distancia no considera los aspectos comportamentales referidos a las medidas (atributos y propiedades contextuales). Estos aspectos en particular son abordados en la siguiente sección.

4.2 Similitud Comportamental

El hecho de compartir atributos o propiedades contextuales entre entidades o contextos permite conocer aspectos comunes entre la definición de proyectos de medición. Sin embargo, esto no dice nada respecto si el comportamiento es similar o no entre dos entidades o contextos.

Cuando no existe conocimiento previo, se podrían obtener estimaciones usando las medidas recolectadas desde sensores IoC, cuantificando las respectivas métricas relacionadas a un atributo o propiedad de contexto. Sin embargo, cuando existe conocimientos previos, estos podrían contrastarse con las estimaciones para analizar si un cambio ha ocurrido. Por ejemplo, un acercamiento basado en Z-scores podría emplearse para analizar comparativamente las medias de dos distribuciones, aunque no sobre las distribuciones de datos en sí mismo [133], [134].

Por un lado, sería posible obtener un conjunto de medidas desde los sensores para dos proyectos dados. Por otro lado, la perspectiva comportamental podría aproximarse contrastando las distribuciones de datos desde los sensores. Así, Sean P y Q las mencionadas distribuciones para dos proyectos, es posible emplear la divergencia de Hellinger (es decir, *H(P, Q)*) para comparar las distribuciones como se expone en la Ecuación 7. Un aspecto importante de mencionar es que la divergencia de Hellinger satisface los axiomas de simetría, no negatividad, identidad en forma conjunta con la desigualdad triangular [135].

Ecuación 7 Divergencia de Hellinger

$$H(P,Q) = \frac{1}{\sqrt{2}} * \sqrt{\sum_{i=1}^{k} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2}$$

Donde p_i y q_i representan la probabilidad de ocurrencia para cada valor de la variable aleatoria, discretizada en *k* intervalos de igual tamaño, promoviendo el cálculo incremental e implementación en hardware de recursos limitados. Sin embargo, los métodos de discretización alternativos representan un tópico abierto y pueden profundizarse como complemento futuro del presente trabajo [136]. De este modo, dado un atributo a_i compartido entre dos entidades (ejemplo, la frecuencia cardiaca), la similitud comportamental se calcula como se indica en la Ecuación 8:

Ecuación 8 Similitud Comportamental para un Atributo

$$\forall a_i \epsilon\ (e_1 \cap e_2): \exists \overrightarrow{X^k}_{e_1,a_i}, \overrightarrow{X^k}_{e_2,a_i} \in\ \mathbb{R} / \ Z(a_i) = H(\overrightarrow{X^k}_{e_1,a_i}, \overrightarrow{X^k}_{e_2,a_i})$$

Donde:

- a_i representa un atributo que pertenece a los atributos comunes entre dos entidades e_1 y e_2.
- $\overrightarrow{X^k}_{e_1,a_i}$ representa la distribución de probabilidad para las medidas del atributo ´a_i´ de una entidad (es decir e_1 o e_2 respectivamente), la cual fue discretizada en ´k´ intervalos de igual magnitud. “k” podría ser arbitrariamente establecido.

La Ecuación 9 se plantea para traducir la Ecuación 8 (es decir, $Z(a_i)$) a un acercamiento orientado a la distancia. De este modo, la idea es que cuando dos conceptos sean similares el valor arrojado se aproxime a cero, mientras que cuando ellos difieren, su valor se acerque a 1.

Ecuación 9 Distancia Comportamental para un Atributo

$$\forall a_i \epsilon\ (e_1 \cap e_2): \exists\ Z(a_i) \in \mathbb{R} \wedge iz(Z(a_i)) \in [0;1] / \ iz\big(Z(a_i)\big) = 1 - Z(a_i)$$

Así, la distancia comportamental interna para dos entidades se calcula de acuerdo con la Ecuación 10.

Ecuación 10 Distancia Comportamental Interna para dos Entidades

$$\forall a_i \epsilon\ (e_1 \cap e_2): idist_{beh} \in [0;1] / \ idist_{beh}(e_1, e_2) = 1 - \frac{\sum_{i=1}^{|e_1 \cap e_2|} iz(Z(a_i))}{|e_1 \cap e_2|}$$

Es importante mencionar que, dado que los estados de entidad se definen a partir de los valores que asumen los atributos de las propias entidades, la distancia comportamental contempla tanto la entidad como sus estados.

Con un razonamiento análogo a las entidades y empleando la distancia de Hellinger (Ver Ecuación 7, la similitud comportamental para los contextos y escenarios se calcula de acuerdo con la Ecuación 11.

Ecuación 11 Similitud Comportamental para dos Contextos

$$\forall p_i \epsilon\ (c_1 \cap c_2): \exists \overrightarrow{X^k}_{c_1,p_i}, \overrightarrow{X^k}_{c_2,p_i} \in \mathbb{R} / Z(p_i) = H(\overrightarrow{X^k}_{c_1,p_i}, \overrightarrow{X^k}_{c_2,p_i})$$

Donde:

- p_i representa una propiedad contextual que comparten dos contextos c_1 y c_2.
- $\overrightarrow{X^k}_{c_1,p_i}$ representa la distribución de probabilidad para las medidas de la propiedad de contexto ´p_i´ de un contexto (es decir c_1 o c_2 respectivamente), la cual fue discretizada en ´k´ intervalos de igual magnitud. “k” podría ser arbitrariamente establecido.

Con un razonamiento análogo al introducido para la Ecuación 9, se pretende llevar a $Z(p_i)$ a una interpretación orientada a la distancia en la Ecuación 12 .

Ecuación 12 Distancia Comportamental para una Propiedad de Contexto

$$\forall p_i \epsilon\ (c_1 \cap c_2): \exists\ Z(p_i) \in \mathbb{R} \wedge ez(Z(p_i)) \in [0;1] / ez\big(Z(p_i)\big) = 1 - Z(p_i)$$

De este modo, la distancia comportamental externa para dos contextos se obtiene empleando la Ecuación 13 como sigue.

Ecuación 13 Distancia Comportamental Externa

$$\forall p_i \epsilon\ (c_1 \cap c_2): edist_{beh} \in [0;1] / edist_{beh}(c_1, c_2) = 1 - \frac{\sum_{i=1}^{|c_1 \cap c_2|} ez(Z(p_i))}{|c_1 \cap c_2|}$$

Al igual que ocurre con los estados de entidad y las entidades, los escenarios se definen a partir de los valores que asumen las propiedades de contexto que los caracterizan. Por tal motivo, la distancia comportamental externa contempla tanto los contextos como sus escenarios asociados.

4.3 Impacto de Escenarios y Estados de Entidad. Una Perspectiva Integrada

Ahora bien, de momento se tiene una perspectiva estructural que permite contrastar la similitud entre definiciones de proyectos contemplando estados de entidad y escenarios. Por otro lado, se cuenta con una perspectiva comportamental que permite contrastar las distribuciones de datos asociadas con atributos de entidad y propiedades de contexto a los efectos de determinar si se desarrollan o no en forma similar. Sin embargo, ellos deben ser analizados en forma conjunta y no aislada, dado que tanto la estructura como su comportamiento permiten conjuntamente determinar cuan similares son dos proyectos de medición.

Para ello, empleando la Ecuación 3 asociada con la distancia estructural interna y la Ecuación 10 relacionada con la distancia comportamental interna, es posible definir el concepto de distancia interna contemplando la estructura del proyecto de medición como así también comportamiento manifestado por los atributos de entidades (y estados asociados), tal y como se expresa la Ecuación 14.

Ecuación 14 Distancia Interna

$$idist(\alpha, \beta, e_1, e_2) = \frac{idist_{str}(\alpha, \beta, e_1, e_2) + idist_{beh}(e_1, e_2)}{2}$$

De igual modo, empleando la Ecuación 6 asociada con la distancia estructural externa junto con la Ecuación 13 relacionada con la distancia comportamental externa, es posible definir la distancia externa de un proyecto de medición contemplando tanto su estructura como su comportamiento, tal y como expone la Ecuación 15.

Ecuación 15 Distancia Externa

$$edist(\gamma, \delta, c_1, c_2) = \frac{edist_{str}(\gamma, \delta, c_1, c_2) + edist_{beh}(c_1, c_2)}{2}$$

Como puede apreciarse en la Ecuación 14 y Ecuación 15, la estructura y comportamiento en el cómputo de las distancias son igualmente ponderadas. Es decir, ambos valores se suman y promedian sin dar mayor importancia relativa a uno u otro. De ser requerido, ambas fórmulas podrían modificarse cuando se desee dar alguna ponderación específica a uno u otro.

En este punto, el interés es aproximar una distancia conjunta o compuesta, es decir, considerando la distancia interna relacionada con la entidad y sus estados (estructura y comportamiento) más la distancia externa contemplando el contexto y sus escenarios. Así, se plantea en la Ecuación 16 la definición para la distancia compuesta capitalizando las definiciones previas.

Ecuación 16 Distancia Compuesta

$$cdist(w, \alpha, \beta, \gamma, \delta, e_1, e_2, c_1, c_2) = w * idist(\alpha, \beta, e_1, e_2) + (1 - w) * edist(\gamma, \delta, c_1, c_2)$$

La distancia compuesta puede determinar el peso relativo que le aporta al componente interno o al externo mediante el parámetro w. Así, cuando se desee dar igual importancia a la distancia externa que a la interna, w será establecido en 0.5. Este parámetro puede ser variado arbitrariamente de acuerdo con la necesidad. Por ejemplo, cuando se requiera considerar solo la perspectiva externa, el parámetro w puede definirse como 0 lo que anulará el primer sumando y mantendrá el segundo en la anterior ecuación. En algunas situaciones, puede que no existan medidas para atributos o propiedades de contexto y por tal no es posible obtener medidas de los sensores. En ese punto, la Ecuación 16 seguiría la perspectiva pertinente para mantenerse operativa.

4.4 Patrón de Aplicabilidad

La presente sección aborda una simulación discreta orientada a estimar un patrón de referencia para el tiempo de cálculo de las distancias definidas a partir de definiciones de proyectos basadas en ECINCAMI. De igual modo, se analiza el tamaño requerido para mantener la matriz de similitud de proyectos a los efectos de procesar un listado descendente respecto a la búsqueda de recomendaciones.

La idea subyacente a la simulación discreta es que la distancia compuesta permita priorizar el espacio de búsqueda basado en similitud de proyectos de medición para guiar la búsqueda de recomendaciones. Así, a partir de una lista de proyecto de medición se obtiene una secuencia ordenada de ellos que sirve para guiar la búsqueda de recomendaciones por similitud (estructural y comportamental) basado en la matriz de similitud obtenida.

Las simulaciones se ejecutaron en Mac Book Pro, 16 GB 2133 MHz LPDDR3, con un procesador de 2,9 GHz Intel Core i7, corriendo macOS Big Sur 11.2.2. Se empleó el JDK 1.8.0_271 y Apache Netbeans IDE 12.1.

La librería *composedIndex* incorpora una clase denominada *Sim* bajo el paquete *io.github.mjdivan.composedIndex* conteniendo dos simulaciones. La misma se encuentra libremente accesible bajo el repositorio denominado *composedIndex* en GitHub (https://github.com/mjdivan/composedIndex).

Las simulaciones parten de los siguientes supuestos:

- Cada proyecto de medición se define siguiendo la estrategia GOCAME-ESVI de acuerdo con la ontología ECINCAMI y se expresa a través de BriefPD,
- Los proyectos comparados corresponden a la misma versión de ECINCAMI,

Como limitaciones y alcances, debe indicarse:

- La distancia compuesta se limita a la ontología ECINCAMI,

- La librería prototípica que implementa el cálculo de distancia y las simulaciones constituyen una referencia para estimar un patrón de tiempo requerido para el cálculo como el espacio de memoria requerido,
- Los resultados pueden variar sensiblemente si se ejecutan sobre sistemas operativos y hardware diferente,
- La totalidad de las librerías y material para reproducir el experimento se encuentran accesibles mediante la licencia Apache 2.0 en GitHub.

La primera simulación tuvo por finalidad medir el tiempo de operación individual durante 10 minutos, manteniendo constante el número de proyectos por mensaje en BriefPD. De este modo, la idea residía en estimar un patrón de referencia para el tiempo de creación de la matriz de similitud, el tiempo de cómputo de la distancia compuesta junto con la cantidad de memoria requerida por la matriz. Para medir el tamaño en memoria de la matriz se empleó la librería jamm (versión 0.3.3) mediante agentes de instrumentación en JAVA. La definición de proyecto se intercambió mediante BriefPD empleando su librería asociada. De este modo, se creó un mensaje con 10 proyectos de medición y durante 10 minutos:

1. Una instancia *ComposedIndex* se creó y midió para estimar el tiempo de inicialización de la matriz, como así también se registró las marcas de tiempo (en inglés, timestamps), el tiempo transcurrido, y el número de proyectos por mensaje.
2. Se incorporaron valores en forma aleatoria dentro de las instancias creadas,
3. Se calculó y midió la distancia para todos los elementos en la matriz,
4. Se midió el tamaño final de la matriz,
5. Pasos 1 a 4 se repitieron cíclicamente hasta alcanzar los 10 minutos. Finalizado el tiempo, Todos los registros de medición se almacenan en un archivo para su análisis.

La segunda simulación focalizó en medir el tiempo de operación individual, pero variando el número de proyectos por mensaje entre 11 y 201, afectando directamente la dimensión de la matriz de similitud. De este modo, Se ejecutaron ciclos desde 11 a 201 proyectos con saltos cada 10 proyecto. Por cada ciclo, se ejecutaron las siguientes operaciones:

1. Se crea una instancia de BriefPD para generar las definiciones de proyecto con el número de proyectos indicados por el ciclo,
2. Se crea y mide la instancia *ComposedIndex* para los mensajes de proyectos actuales. Se mide la marca de tiempo, tiempo transcurrido, y el número de proyectos involucrados,
3. Se incorporan valores aleatorios en la matriz y se calcula la distancia para todas las combinaciones,

4. Se mide el tamaño de la matriz para los proyectos involucrados,
5. Pasos 1 a 4 se repitieron hasta culminar con 201 proyectos, luego de lo cual las medidas se almacenan en un archivo para su análisis.

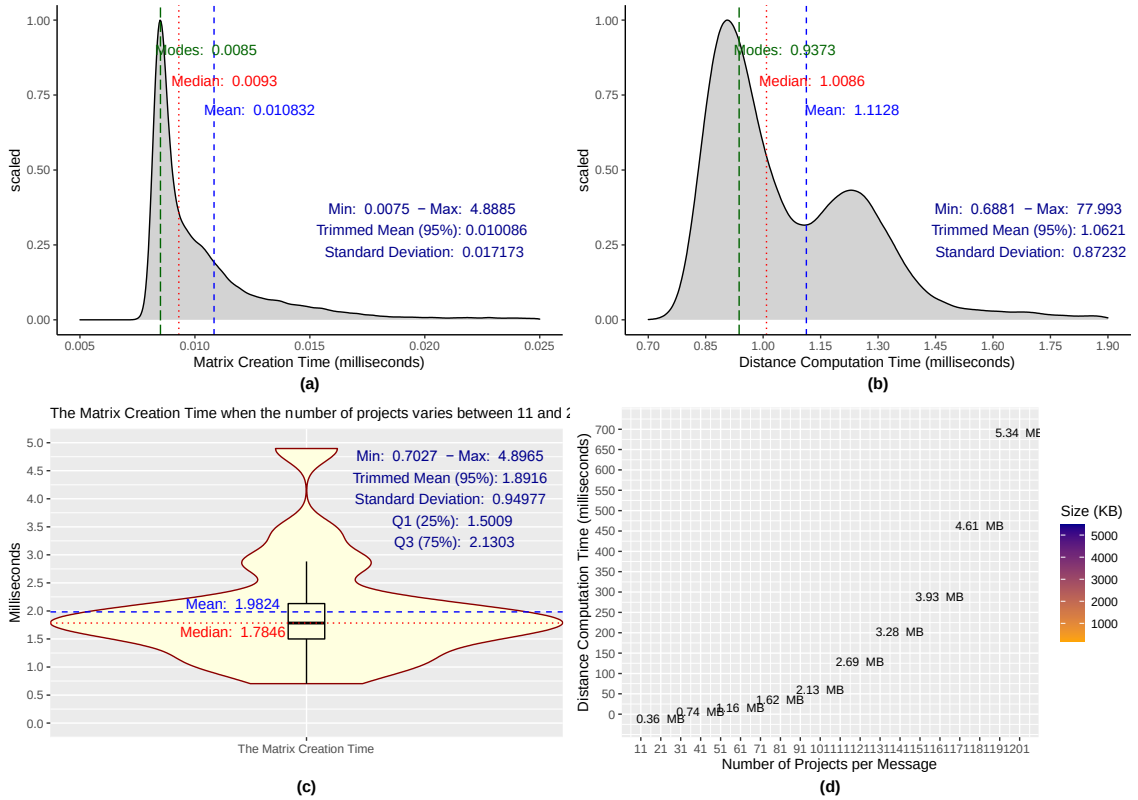

Figura 18. a) Curva de densidad para el tiempo de creación de la matriz (Simulación 1); b) Curva de Densidad para el tiempo de cómputo de la distancia compuesta (Simulación 1); c) Gráfico de violín del tiempo de creación de la matriz cuando el número de proyectos por mensaje varía entre 11 y 201 (Simulación 2); d) Evolución del tamaño de la matriz y el tiempo de cómputo de la distancia cuando el número de proyectos varía entre 11 y 201 (Simulación 2).

La Figura 18.a muestra la curva de densidad obtenida para el tiempo individual de creación de la matriz, alcanzando 0.01 ms para crear la matriz a partir de la definición de proyectos con diez proyectos. Los valores atípicos (o outliers) en este contexto se asocian con la interferencia del recolector de basura de JAVA. Como se puede apreciar tanto la moda como la mediana tiene un valor inferior que la media recortada (en inglés, trimmed mean), lo que sugeriría una perspectiva conservadora de los tiempos obtenidos.

La Figura 18.c describe la misma variable (tiempo individual de creación de la matriz) pero variando el número de proyectos entre 11 y 201 por mensaje (Simulación 2), alcanzando una mediana para la creación de la matriz de 1.78 ms. Como una diferencia respecto de la Figura 18.a (Simulación 1), la simulación 2 creó los proyectos dentro de cada ciclo y ello incorporó una sobrecarga en el tiempo.

La Figura 18.b describe la curva de densidad para el tiempo de cálculo de la distancia, alcanzando 1.06 ms como media recortada (95%) para ejecutar la totalidad de los

cálculos, utilizando un mensaje con 10 definiciones de proyectos (Simulación 1). En este sentido, vale la pena mencionar que cada elemento de la matriz no es un elemento simple sino compuesto, contenido los resultados parciales de las ecuaciones aquí introducidas (Ver la clase *ComposedSimilarityNode* bajo el paquete io.github.mjdivan.*composedIndex*). Esto permite decidir el tipo de distancia a utilizar para filtrar proyectos o recomendaciones, por ejemplo, sería posible filtrar basado solo en la distancia interna.

La Figura 18.d expone la evolución del tiempo de cómputo de distancia y tamaños cuando el número de proyectos varía entre 11 y 201, afectando la dimensionalidad de la matriz. La implementación aquí propuesta empleó una matriz triangular como un arreglo, incorporando mapeo unidimensional para obtener mejores resultados. La totalidad de las distancias para una matriz triangular con 201 proyectos podría ser almacenada en 5.34 MB consumiendo solo 716 ms para desarrollar la totalidad de los cálculos. Si bien se trata de una prueba de concepto plausible de mejora de cuerdo al contexto de aplicación y plataforma, estos resultados proveen una referencia en tiempo y tamaño requerido para su uso.

4.5 Conclusiones Generales del Capítulo

El presente capítulo introdujo la distancia interna y externa basado en los conceptos, términos, y relaciones definidas en la ontología de medición y evaluación ECINCAMI. Las distancias analizadas contrastan la perspectiva estructural y comportamental. Desde el punto de vista estructural, se define el modo en que los conceptos se relacionan y cuantifican. Esta estructura proviene de la definición del proyecto y se expresa mediante BriefPD. Desde el punto de vista comportamental, se emplea las medidas relacionadas con las métricas del proyecto para analizar los cambios y evolución de los conceptos monitoreados.

Dado que, independientemente de la definición del proyecto de medición, la entidad y su entorno podrían cambiar a lo largo del tiempo, la distancia compuesta contempla tanto los estados de entidad como los escenarios relacionados con el contexto. Un cálculo basado en la idea de Z-score permite obtener un valor numérico que expresa cuan similares (o diferentes) dos proyectos son. Se emplea la divergencia de Hellinger para poder comparar las distribuciones de datos (medidas) para las diferentes características analizadas. En este sentido, es importante destacar que La divergencia de Hellinger satisface los axiomas de simetría, no negatividad, identidad en forma conjunta con la desigualdad triangular. La distancia compuesta permite articular tanto la distancia interna como la externa incorporando un parámetro de ponderación que puede ser arbitrariamente definido.

De este modo, la distancia compuesta permite ordenar por similitud un conjunto de proyectos de medición basados en ECINCAMI, contemplando su definición y el comportamiento manifestado por sus diferentes variables (es decir, las métricas que

cuantifican los atributos y propiedades de contexto). A partir de lista ordenada de proyectos, es posible obtener de ellos una serie de criterios de decisión, recomendaciones, y clasificadores lo más similar posible a una nueva situación (desconocida o sin historia) de otro proyecto.

Se llevaron a cabo dos simulaciones discretas para estimar o contar con un patrón de referencia respecto al tamaño que requeriría en memoria tener la información de similitud desagregada, como así también su costo de procesamiento (tiempo). A través de estas, se pudo obtener un tiempo de cómputo de la distancia de 1.06 ms para toda la matriz con 10 proyectos, mientras que el tiempo de creación de esta fue 0.01 ms y representó 0.36 MB. Los resultados son alentadores en áreas donde se requiere procesamiento de datos en tiempo real para filtrar un conjunto de recomendaciones previo a efectuar el plan de consulta. Esto es, de qué modo poder ir a buscar información potencialmente pertinente específicamente a proyectos similares en una memoria organizacional antes de realizar la consulta en sí, guiado por su situación actual (comportamiento) y definición (estructura).

A su vez, la matriz de similitud contiene la distancia como un elemento compuesto, donde se describe las diferentes instancias de cálculo intermedio para su obtención. De esta manera, se permite un filtrado selectivo de proyectos basado en diferentes niveles de agregación y regulado por cualquier combinación arbitraria de los parámetros α (importancia relativa de los estados y sus transiciones), β (Importancia relativa de la entidad respecto de sus estados), γ (importancia relativa de los escenarios respecto de sus transiciones), δ (importancia relativa del contexto respecto de sus escenarios), o w (importancia relativa de la distancia interna respecto de la externa).

Hasta aquí se ha abordado la importancia e intercambio del proyecto de medición basado en ECINCAMI y mediante BriefPD junto con el modo en que pueden compararse los mismos basado en una perspectiva estructural y comportamental para guiar la búsqueda de recomendaciones potencialmente similares cuando no existe conocimiento previo. El siguiente capítulo introduce la arquitectura de procesamiento de flujos de datos basado en metadatos de mediciones organizado por capas y niveles. Allí confluye la definición de proyecto como elemento para formalizar la captación, procesamiento, intercambio, y análisis de datos, como así también, el rol del clasificador incremental para tomar decisiones y la distancia compuesta para buscar recomendaciones que las sustenten.

Capítulo 5

Arquitectura de Procesamiento basada en Metadatos de Mediciones

Capítulo 5. Arquitectura de Procesamiento basada en Metadatos de Mediciones

Introducción

El capítulo anterior ha introducido la distancia compuesta para calcular la similitud estructural y comportamental de dos o más proyectos de medición, definidos de acuerdo con la ontología ECINCAMI e intercambiados utilizando el formato BriefPD. Ello permite ordenar los proyectos similares en forma descendente para localizar clasificadores alternativos cuando un proyecto dado no cuenta con suficiente información para hacerlo.

De este modo, es posible el intercambio de definiciones de proyectos de medición entre diferentes componentes del sistema de recolección, como así también determinar en qué medida los proyectos intercambiados son similares entre ellos considerando su definición y comportamiento.

El concepto de estrategia de monitoreo de proyectos de medición basada en metadatos de medición fue introducido en [137] basado en el marco C-INCAMI original, mientras que sus procesos fueron especificados en [138], [139]. Sin embargo, a partir de la extensión y obtención de ECINCAMI, la estrategia extendida de soporte GOCAMI-ESVI y funcionalidades añadidas es que surge la Arquitectura de Procesamiento basado en Metadatos de Medición (en inglés PAbMM) como una evolución natural de aquella.

Sintéticamente, PAbMM organiza la recolección de datos a nivel de componentes y procesamiento basado en la definición de proyecto de medición mediante BriefPD. Al inicio, cada recolector de datos empareja sus sensores respecto de la métrica que instrumentará. En tal sentido, es importante mencionar que la métrica puede vincularse a un atributo de la entidad o una propiedad del contexto. Una vez que el recolector define el emparejamiento entre sensor (o instrumento) y la métrica que implementará, este comienza a recolectar las medidas y las almacena en un buffer local de acuerdo con la política de transmisión que defina (por ejemplo, esto es importante para la optimización del tiempo de vida de las baterías). Al momento en que el colector informa las medidas, lo hace en forma conjunta con etiquetas (metadatos) que describen su pertenencia conceptual. Es decir, se sabe que un número dado (ejemplo, 37) proviene de un sensor de temperatura ambiental que implementa una métrica dada para el contexto de una entidad en particular. Ese flujo de datos y metadatos se intercambia en un formato específico entre recolectores (cuando existe transmisión intermedia) y-o directamente a las capas de procesamiento principales. Las capas de procesamiento reciben el flujo de medidas y metadatos de todos los recolectores y las organizan, leen, procesan, analizan, interpretan y toman decisiones guiados por los metadatos. El cómputo de escenarios y estados actuales de la entidad se lleva adelante en línea consumiendo los metadatos y sus valores asociados, lo que permite orientar la toma de decisiones para una entidad y contexto dado. Tomada una decisión, la experiencia

previa y conocimiento permite soportar recomendaciones. De no existir información previa, la búsqueda de estas es guiada mediante la lista de similitud de proyectos obtenida mediante la distancia compuesta.

Este capítulo introduce la vista arquitectural del procesamiento basado en metadatos de medición como un todo para tener la perspectiva global. Se analizarán los detectores de cambios de datos incrementales junto con los mecanismos de descarte selectivos basados en metadatos de medición. Adicionalmente, se detallará el modo en que escenarios y estados de entidad se calculan en tiempo real a partir del procesamiento del fujo compuesto de medidas (es decir, datos más metadatos). Se aborda las funcionalidades por capa y nivel de servicio junto con las estrategias de búsqueda.

El capítulo se organiza en siete secciones. La primera sección describe en términos generales la arquitectura de procesamiento de flujos de datos. La segunda sección aborda la naturaleza distribuida de la recolección de datos y su vínculo con la definición del proyecto de medición. La tercera sección describe el procedimiento de determinación de estados y escenarios, como así también de las probabilidades de transición entre ellos a partir del flujo de medidas. La cuarta sección introduce el proceso de reunión de flujos de medidas considerando todas las pasarelas y adaptadores de medición. La quinta sección aborda el tipo de análisis que se realiza sobre las medidas recolectadas. La sexta sección introduce el rol de la toma de decisiones, las recomendaciones y su vínculo con la distancia compuesta. Finalmente, se describen las conclusiones del capítulo.

El capítulo se soporta en las siguientes publicaciones efectuadas a lo largo del proceso de investigación:

- Diván, M. Sánchez-Reynoso, M. Gonnet, S. (2022) "**Measurement Project Interoperability for Real-time Data Gathering Systems**". Future Generation Computer Systems, Elsevier, ISSN 0167-739X, 129, 298-314 https://doi.org/10.1016/j.future.2021.11.031

- Diván, M. Sánchez-Reynoso, (2021). "Big Data Analysis for Green Computing: Concepts and Applications". **Effect of the measurement on Big Data Analytics. An evolutive perspective with Business Intelligence**. ISBN 9781003032328. pp 50-69. Estados Unidos. CRC Press. http://dx.doi.org/10.1201/9781003032328-4

- Diván, M. Sánchez Reynoso, M. (2021) "**A Metadata and Z Score-based Load-Shedding Technique in IoT-based Data Collection Systems**". International Journal of Mathematical, Engineering and Management Sciences. ISSN: 2455-7749. e-ISSN: 2455-7749. Elsevier. Vol.6, nro 1. pp 363 – 382. https://ijmems.in/volumes/volume6/number1/23-IJMEMS-SBS19-34-6-1-363-382-2021.pdf

- Diván, M & Sánchez Reynoso, M (2020) "**A Real-Time Entity Monitoring based on States and Scenarios**" CLEI Electronic Journal. ISSN 0717-5000. Vol. 23 (1). Pp 2-1:2-25. https://doi.org/10.19153/cleiej.23.1.2

- Diván, M & Sánchez Reynoso, M (2020) "**Optimizing Data Transmission from IoT devices through Weighted Online Data Changing Detectors**" Advances in Data Science and Adaptive Analysis. ISSN 2424-922X. Vol 12 (2). 2041001 (pp.1:33). https://doi.org/10.1142/S2424922X20410016

- Diván, M & Sánchez Reynoso, M (2020) "**Relocating the Load-Shedding Strategy in the Data Stream Processing Architecture**" In IEEE 2020 Argencon. Resistencia, Chaco. 2 al 4 de diciembre. http://dx.doi.org/10.1109/ARGENCON49523.2020.9505446

- Sánchez Reynoso, M &Diván, (2020) "**Recomendación por similitud semántica en repositorios con grandes volúmenes de datos de medición**". V Jornadas de Intercambio y Difusión de los Resultados de Investigaciones de los Doctorandos de Ingeniería. UTN Córdoba, Córdoba. 6 y 7 octubre. ISBN: 978-950-42-0200-4. https://doi.org/10.33414/ajea.5.751.2020

- Diván, M, Sánchez Reynoso, M & Abd Wahab, M (2020) "**Dynamic Switching in the Measurements' Collecting from Heterogeneous Data Sources**". Journal of Physics: Conference Series. ISSN 1742-6596. Vol 1529. 022058:1-8. https://doi.org/10.1088/1742-6596/1529/2/022058

- Sánchez Reynoso, M & Diván, M (2019) "**Improving the Real-Time Searching in the Organizational Memory**". Procedia Computer Science. Elsevier Ltd. Vol. 154, pp. 293-304. ISSN: 1877-0509. https://doi.org/10.1016/j.procs.2019.06.043

- Diván, M & Sánchez Reynoso (2019) "**A Load-Shedding Technique based on the Measurement Project Definition**". In V. Jain, S. Patnaik, F. Popentiu Vladicescu, and I.K. Sethi (Eds.). Proceedings of 5th International Conference on Intelligent Computing, Communication & Devices (ICCD 2018), Xi'an, China, November 22-24 of 2018. In Advances in Intelligent Systems and Computing, Springer Nature Singapore. pp.1027-1033. ISSN 2194-5357. https://doi.org/10.1007/978-981-13-9406-5_122

- Diván, M & Sánchez Reynoso, M (2019) "**Extending the Data Stream Processing Strategy to Scenario Analysis**". International Journal of Advanced Trends in

Computer Science and Engineering. ISSN: 2278-3091. Vol. 8(1.4):1-8. http://dx.doi.org/10.30534/ijatcse/2019/0181.42019

- Diván, M & Sánchez Reynoso, M (2019) "**An Architecture for the Real-Time Data Stream Monitoring in IoT**". Book Chapter in "Multimedia Big Data Computing for IoT Applications: Concepts, Paradigms, and Solutions", S. Tanwar, S.Tyagi, and N. Kumar (Eds.). pp. 59-100. ISBN 978-981-13-8759-3, Springer. https://doi.org/10.1007/978-981-13-8759-3_3

- Diván, M & Sánchez Reynoso, M (2018) "**The Real-Time Measurement and Evaluation as System Reliability Driver**". Book Chapter in "System Reliability Management: Solutions and Technologies". Anand, A & Ram, M (Eds.). CRC Press, Taylor & Francis Group. Pp. 161-188. https://doi.org/10.1201/9781351117661-11

5.1 Una Perspectiva Global de la Arquitectura de Procesamiento

Los dispositivos asociados con Internet de las Cosas (IoC) suelen ser alternativas económicas y accesibles para implementar diferentes estrategias de monitoreo. Por esa razón, no suena extraño disponer de diferentes dispositivos midiendo (o monitoreando) distintos aspectos relacionados con el entorno, hogar, entre otras aplicaciones [140].

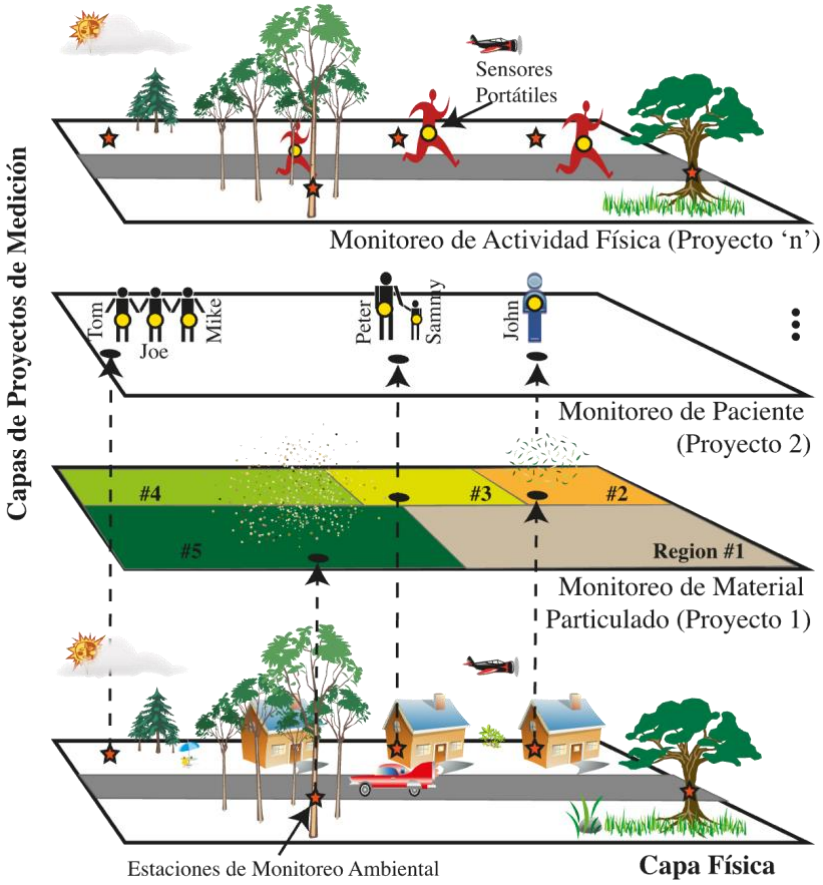

Figura 19 Visión Transversal de los Proyectos de Medición

La idea de reutilizar dispositivos instalados previamente (aún cuando no hayan tenido el mismo objetivo inicial) suena interesante para optimizar un presupuesto.

Incluso podría permitir contar con puntos de vistas adicionales, acceder a datos históricos (en caso de disponibilidad) y compartirse entre varios proyectos de medición, tal y como puede apreciarse en la Figura 19. Sin embargo, ello requeriría una articulación entre las fuentes de datos que son heterogéneas en términos del instrumento (o dispositivo) pero también en base al proyecto en el cual se incorporaron originalmente. Por un lado, los dispositivos heterogéneos debieran articularse consistentemente con otros proyectos de medición (en [141], se menciona como puente semántico). Por otro lado, se plantea una relación complementaria entre proyectos de medición.

Cada proyecto de medición representa un diseño experimental alineado con un objetivo definido por el usuario. Esto es una perspectiva transversal para describir cómo los datos (o medidas) son recolectadas. Por ejemplo, la Figura 19 describe una capa física donde existen un conjunto de estaciones de monitoreo ambiental preinstaladas

(indicadas con estrellas). A su vez, existen tres proyectos de medición diferentes por capa: (1) El proyecto 1 se focaliza en el monitoreo de material particulado, (2) El proyecto 2 se centra en el monitoreo del paciente ambulatorio, (3) El proyecto 3 aborda el monitoreo de la actividad física. Claramente, ninguno de ellos comparte su objetivo y focaliza en características y entidades diferentes. Ahora bien, ello no implica que un proyecto no pueda capitalizar características que le pueden ser interesantes y complementarias (aún con divergencia del objetivo). Por ejemplo, el proyecto 1 podría reutilizar información de las estaciones de monitoreo ambiental para complementar las lecturas de material particulado (por ejemplo, humedad y temperatura). Similarmente, el proyecto 2 podría aprovechar las estaciones de monitoreo ambiental y las lecturas de material particulado para caracterizar las zonas por donde un paciente ambulatorio está caminando. El capítulo 3 ha introducido el rol de la estrategia GOCAME-ESVI para describir el proyecto de medición basado en la ontología ECINCAMI e intercambiarlo mediante BriefPD (alternativamente JSON o XML).

Ahora bien, una vez que cada proyecto se define y cuenta con el respectivo contenido intercambiable mediante BriefPD, es necesario recolectar, procesar, y analizar las medidas guiadas por tal definición para poder implementar el monitoreo. En este punto es donde toma especial interés la Arquitectura de Procesamiento de Datos basada en Metadatos de Mediciones (en inglés PAbMM) [117], [142] introducida en la Figura 20.

La arquitectura PAbMM se encuentra organizada para satisfacer dos perspectivas de procesamiento alrededor de un proceso de medición: la consolidación central y recolección distribuida.

Por un lado, el procesamiento de consolidación central se basa en la nube para incorporar confiabilidad y escalabilidad. Posee una organización multinivel caracterizada como sigue:

- Organizarse en capas. Cada una se asocia con un servicio requerido para la automatización de la medición y recomendación (es decir, gestión de dispositivos, diseño experimental, gestión de conocimiento, recolección de datos, analítica y servicios de datos).

- Cada capa actúa en forma autónoma para promover la paralelización, aunque ellas se encuentran relacionadas unas con otras para servir el proceso de medición.

Por otro lado, el monitoreo en campo implica áreas amplias de cobertura y cierto nivel de resolución (es decir, dispositivos por área). Tanto la nube como la cobertura en campo son acotados en alcance por presupuesto y la tecnología de comunicación. De este modo, el acercamiento de recolección se organiza jerárquicamente alrededor de pasarelas y adaptadores de medición (o puentes semánticos). Las pasarelas cuentan con una mejor configuración de hardware y soportan memorias caché entre la nube y adaptadores de medición. Los adaptadores de medición (o puentes semánticos) tratan

directamente con los sensores. Ellos traducen datos planos desde el sensor en un formato específico siguiendo la definición del proyecto (es decir, BriefPD). Una pasarela puede coordinar uno o más puentes semánticos, haciendo escalable el monitoreo a través de la jerarquía.

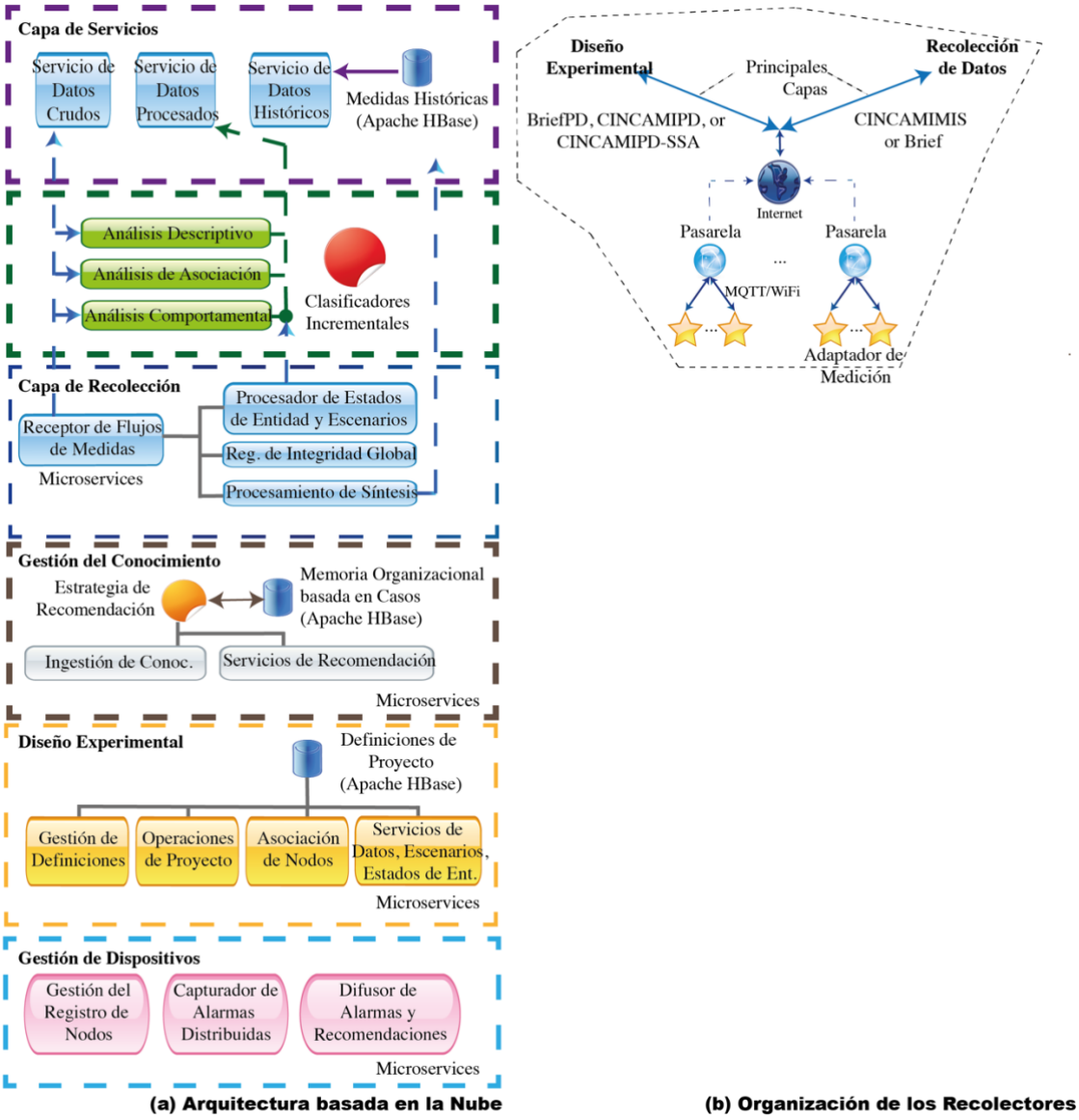

Figura 20 Perspectiva General de la Arquitectura de Procesamiento basada en Metadatos de Mediciones

PAbMM permite automatizar el proceso de medición utilizando la definición del proyecto mediante BriefPD. El contenido se carga en memoria de acuerdo con el modelo de objetos (Ver ECINCAMI en Capítulo 3) y se crean las estructuras de datos en memoria para interpretar y procesar los datos (medidas) en tiempo real [139]. Esto constituye el paso inicial en la arquitectura para cualquier proyecto de medición. De igual modo, los adaptadores de medición emplean el contenido de BriefPD para identificar las entidades que monitorean y cómo emparejar sus sensores respecto de las métricas que implementan. Es decir, procesado el archivo de definición del proyecto, cada adaptador

de medición sabe que un sensor dado provee valores para una métrica determinada y que esta cuantifica un atributo o propiedad contextual de una entidad o su contexto.

Por ello, los dispositivos en campo (sean pasarelas o adaptadores de medición) reciben el conjunto de definiciones con quienes ellos colaboran. De este modo, (1) Los dispositivos conocen cómo traducir un dato crudo desde los sensores en flujos de medidas con etiquetas a través del adaptador de medición (o puente semántico) [141]–[143] y (2) PAbMM conoce cómo interpretar cada etiqueta, ID, entre otros elementos empleando la definición del proyecto (BriefPD).

Desde la perspectiva basada en la nube, las capas de PAbMM tienen los siguientes objetivos:

- **Gestión de Dispositivos**: Es responsable de mantener actualizado un repositorio central con los dispositivos involucrados en la estrategia de recolección de datos para todos los proyectos de medición activos, y, además, de mantener la comunicación de datos bidireccional con ellos (por ejemplo, para enviar alarmas basado en los datos analizados).

- **Diseño Experimental**: Es responsable de gestionar cada definición de proyecto de medición (activa o no), un registro con el conjunto de operaciones asociadas, los nodos vinculados (adaptador de medición o pasarela), y de mantener actualizado la probabilidad para escenarios y estados de entidad de acuerdo con el procesamiento de las medidas recibidas. Esta capa es quien inicializa un proyecto de medición, mientras que la capa de gestión de dispositivos comunica mediante BriefPD los distintos proyectos a los respectivos nodos.

- **Gestión del Conocimiento**: Su función esencial es la gestión de experiencias previas y de conocimiento específico para cada proyecto. Se encuentra organizada en base a casos, donde cada atributo (o propiedad contextual) representan una característica. De este modo, los clasificadores incrementales emplean dicho conjunto de datos para soportar un razonamiento basado en casos dentro y entre proyectos. Aquí es donde se emplea la distancia compuesta introducida en el capítulo 4.

- **Recolección o Reunión de Datos**: Como su nombre representa, esta capa se focaliza en recibir el flujo de medidas (datos y metadatos) desde los adaptadores de medición, procesarlos e interpretarlos siguiendo la definición del proyecto. Adicionalmente, actualiza las probabilidades relacionadas con los estados de entidad y escenarios (articulado con la capa de diseño experimental) y crea una síntesis de datos que se almacena en la base de datos columnar [144]. Como registro de integridad, un árbol de Merkle es continuamente actualizado de acuerdo con las medidas recibidas de cada dispositivo.

- **Analíticas**: Esta capa analiza los datos en tiempo real para actualizar la estadística descriptiva por métrica de proyecto, ejecutar el análisis de asociación entre métricas (ejemplo, análisis de correlación), y el análisis comportamental desde la perspectiva de la distribución de datos.

- **Servicios de Datos**: Esta capa facilita el consumo de datos a terceros bajo modalidad de suscripción en tres modos. El servicio de datos crudos replica el flujo de medidas etiquetado (datos y metadatos) por proyecto tal y como arriba desde las fuentes de datos. El servicio de datos procesados provee acceso bajo demanda a los resultados del procesamiento por proyecto (ejemplo, estadística descriptiva). Finalmente, el servicio de datos históricos provee acceso a los datos procesados por proyecto de acuerdo con la política de síntesis.

Cada capa se implementa mediante microservicios y tiene un funcionamiento autónomo que le permite funcionar parcialmente (o con degrades en la calidad del servicio) cuando otra capa se torne no disponible. Por ejemplo, si se cortare la recolección de datos, el servicio de datos seguiría funcionando con la última información conocida.

Desde la perspectiva de Internet de las Cosas, se han introducido dos tipos de roles en los dispositivos: Pasarela y Adaptador de Medición (o Puente Semántico). Las pasarelas y adaptadores de medición se organizan jerárquicamente y mantienen el registro unificado de nodos utilizando una base de datos distribuida basada en Blockchain [145] (Ver Capítulo 6). Sus principales funciones son:

- **Pasarelas**: Es responsable de proveer servicios de caché para las definiciones de proyecto y recomendaciones entre la nube y adaptadores de medición. A su vez, soporta las transmisiones de datos indirectas (Ver sección 5.2.3) desde los adaptadores de medición ante situaciones particulares (ejemplo, el adaptador se encuentra fuera de alcance para una transmisión directa).

- **Adaptador de Medición**: Actúa como puente semántico guiado por la definición del proyecto de medición. Es decir, toma los datos crudos de uno o más sensores y los traduce a flujos etiquetados incorporando metadatos que describen su semántica (por ejemplo, la métrica a la que pertenece cada número o distribución de probabilidad). El flujo de medición etiquetado es transmitido (directamente o no) hacia la capa de recolección de datos mediante microservicios.

Dado que un dispositivo es autónomo y puede ser reutilizado entre proyectos, la relación entre dispositivos es direccionada mediante una base de datos basada en Blockchain. Es decir, incluso cuando se cuenta con un registro central en la nube (Ver

capa gestión de dispositivos en la Figura 20, el inventario de dispositivos se gestiona en forma distribuida en campo. De este modo, cada dispositivo (sea una pasarela o adaptador de medición) es el único autorizado para registrar/actualizar/borrar sus datos descriptivos en la cadena de bloques (ejemplo, dirección IP, clave pública, puertos disponibles, servicios de datos, etc.). El empleo de esta tecnología permite asegurar la trazabilidad de cada cambio incorporado en los registros, proveyendo confiabilidad sobre la información de contacto de un dispositivo. En el capítulo seis se volverá sobre este aspecto y se proveerán detalles.

5.2 Recolección de Datos Distribuida y Definición de Proyecto de Medición

Hasta este momento, se ha descrito el rol de la definición de proyectos en PAbMM y cómo permite en los adaptadores de medición asociar métricas y sensores. Sin embargo, no se ha dado detalles sobre el esquema de intercambio de mediciones, es decir, aquel que permite informar datos (medidas) junto con metadatos a la capa de recolección de datos. La sección 5.2.1 introduce y describe el esquema de intercambio de medidas a los efectos de visualizar el impacto del metadato en el procesamiento del flujo de medidas. La sección 5.2.2 detalla la articulación entre adaptadores de medición y pasarelas en la recolección de datos distribuida. Finalmente, la sección 5.2.3 describe la estrategia para la transmisión indirecta de datos.

5.2.1 Esquema de Intercambio de Mediciones

A partir de la ontología ECINCAMI, se ha organizado la jerarquía de conceptos para el intercambio de medidas denominado CINCAMI/MIS (acrónimo en inglés para Measurement Interchange Schema) como muestra la Figura 21.

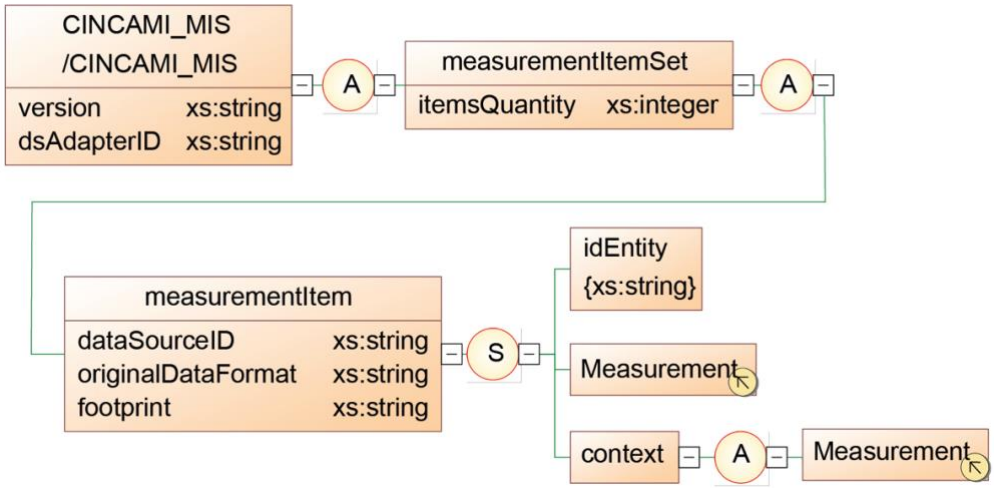

Figura 21 Nivel Superior del mensaje CINCAMI/MIS

En la organización jerárquica del esquema, es posible encontrar tres tipos de símbolos junto a los conceptos:

- **A**: Representa que es posible contar con un conjunto de otras etiquetas en cualquier orden. Por ejemplo, debajo de la etiqueta *measurementItemSet* es posible encontrar un conjunto de etiquetas *measurementItem* en cualquier orden.

- **S**: Indica que el conjunto de etiquetas que contiene aparecerá en un orden específico. Por ejemplo, bajo la etiqueta *measurementItem*, encontrará las etiquetas *idEntity*, *Measurement*, y *context* en dicho orden.

- **C**: Representa que solo una de las etiquetas que contiene puede ser escogida entre todas las alternativas listadas en la misma.

La Figura 21 representa el nivel superior del esquema de intercambio de mediciones. La etiqueta *CINCAMI_MIS* limita el mensaje para un único adaptador de medición. Esta etiqueta contiene los atributos *version* y *dsAdapterID*. El primero refiere a la versión del esquema utilizado en el mensaje; mientras que el segundo refiere al identificador único del adaptador de medición. Las medidas se agrupan mediante la etiqueta *measurementItemSet*, que contiene un conjunto de etiquetas *MeasurementItem*. Esta última etiqueta representa una medida asociada con información de su contexto (más datos complementarios) y cuenta con los siguientes atributos: 1) Identificador de la fuente de datos o sensor (*dataSourceID*), 2) Formato de datos crudo original relacionado con el sensor, y 3) huella basada en MD5 que permite verificar la integridad del contenido (etiquetas y datos en niveles inferiores). Adicionalmente, se identifica la entidad para la que se asocia la medición (*idEntity*), junto con el detalle de la medida (*Measurement*) y su contexto asociado (*context*).

El círculo amarillo con una flecha en su interior representa la reutilización del mismo concepto, es decir, el contenido de la etiqueta *Measurement* bajo las etiquetas *measurementItem* o *context* es estructuralmente idéntico. La Figura 22 introduce la estructura interna para cada medición que se presenta contraída en la Figura 21.

Bajo la etiqueta *Measurement* se encuentra una secuencia ordenada de etiquetas para describir el instante con la zona horaria de la medición (etiqueta *datetime*), la métrica con la que se asocia la medida (es decir, *idMetric*) e información de la medida en sí (bajo la etiqueta *Measure*). Es importante mencionar que tanto el identificador de entidad (Ver *idEntity* en la Figura 21) como el de la métrica (Ver *idMetric* en la Figura 22) se toman desde la definición del proyecto, motivo por el cual, hasta no procesarse el contenido del archivo BriefPD no es posible recolectar medidas.

La etiqueta *quantitative* permite separar los datos de la medida respecto de los datos complementarios informados con ella (*complementaryData*). La medida cuantitativa puede ser determista o estimada. Si fuere determinista, se informa el valor numérico único en la etiqueta *deterministicValue*. Sin embargo, cuando existe una

distribución de valores con su respectiva probabilidad, ellos son informados bajo la etiqueta *likelihoodDistribution,* como un conjunto de etiquetas *estimated* que describe el par valor y probabilidad (etiquetas *value* y *likelihood* respectivamente).

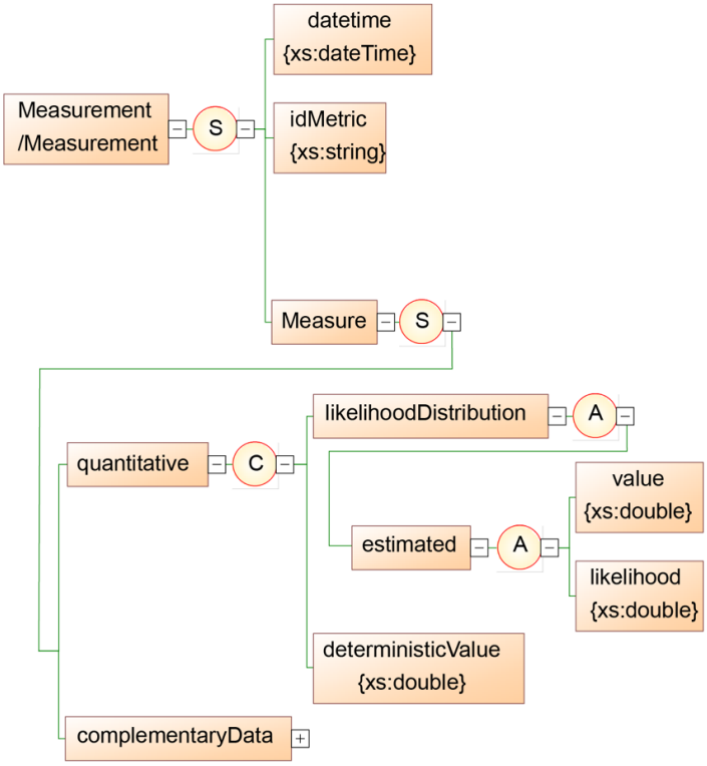

Figura 22 Estructura de Cada Medición en CINCAMI/MIS

La Figura 23 introduce la organización de los datos complementarios de CINCAMI/MIS, extendiendo la etiqueta *complementaryData* de la Figura 22.

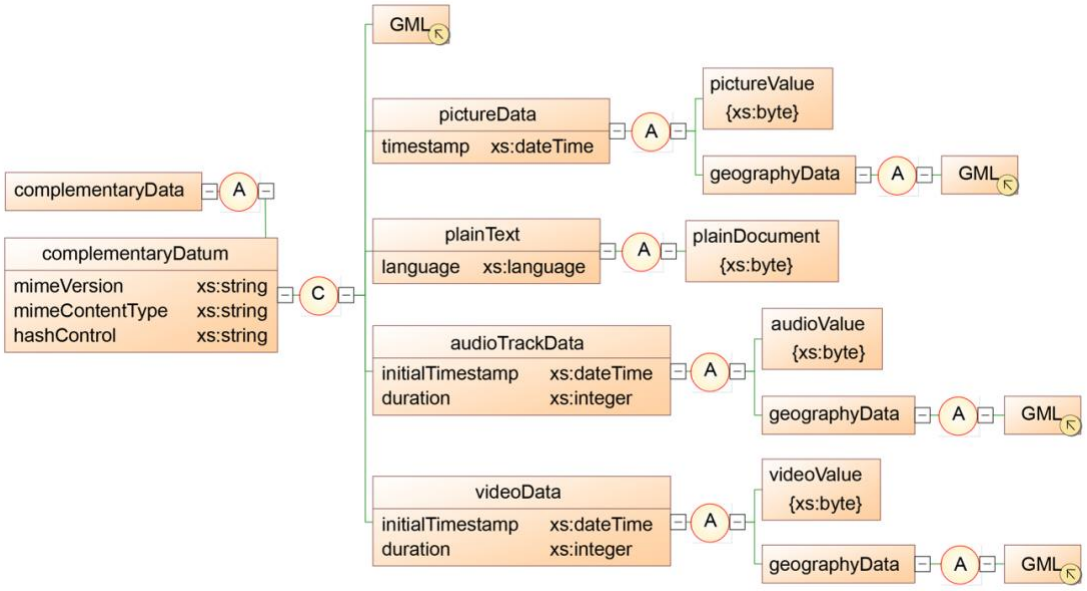

Figura 23 Organización de los Datos Complementarios en CINCAMI/MIS

Una medida puede contener uno o más datos complementarios organizados bajo la etiqueta *complementaryDatum*. Sin embargo, bajo cada etiqueta *complementaryDatum* puede escogerse solo un tipo dato complementario: 1) Un documento organizado bajo el lenguaje de marcado geográfico (en inglés, *Geography Markup Language -GML*); 2) Una imagen describiendo el contexto o alguna característica de la entidad bajo análisis (*pictureData*); 3) Un texto plano describiendo, por ejemplo, un registro de operaciones(*plainText*); 4) Una pista de audio representativa de algún atributo o propiedad de contexto relacionado con la entidad o su propiedad de contexto respectivamente (*audioTrackData*); 5) Una secuencia de video que podría describir algún aspecto interesante de una regiónVideo (*videoData*). Adicionalmente, tanto la secuencia de audio, video, o imagen pueden asociarse con datos geográficos para indicar su posición asociada.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<cincamimis Version="2.0" dsAdapterID="DSA_1">
  <measurementItemSet itemsQuantity="2">
    <measurementItem dataSourceID="ds_temp" originalDataFormat="format"
 2  footprint="c8633391bca18da1767336723ffac536" projectID="PRJ_1"
      entityCategoryID="EC1">
      <idEntity>Peter</idEntity>
      <Measurement>
 3      <datetime>2020-03-24T10:55:46.461-03:00</datetime>
        <idMetric>dm_ctemp</idMetric>
        <Measure>
          <quantitative>
 4          <likelihoodDistribution>
              <Estimated>
                <value>37.36506</value>
                <likelihood>0.33333</likelihood>
              </Estimated>
              <Estimated>
                <value>37.40985</value>
                <likelihood>0.33333</likelihood>
              </Estimated>
              <Estimated>
                <value>37.38850</value>
                <likelihood>0.33333</likelihood>
              </Estimated>
            </likelihoodDistribution>
          </quantitative>
        </Measure>
      </Measurement>
    </measurementItem> ...
```

Figura 24 Vista Parcial de Un Mensaje CINCAMI/MIS organizado mediante XML

La Figura 24 muestra una representación parcial de un mensaje CINCAMI/MIS organizado mediante XML para facilitar su interpretación. Cada etiqueta en el mensaje se sustenta en el esquema de intercambio y por carácter transitivo en la ontología ECINCAMI. De este modo, por ejemplo, el círculo indicado con un uno en su interior, señala el adaptador de medición del cual proviene las medidas (es decir, *DSA_1*) y del instrumento utilizado para obtenerla (es decir, sensor *ds_temp* vinculado al adaptador).

El círculo con un dos en su interior describe: a) La huella calculada para su contenido inferior a los efectos de verificar su integridad, b) El identificador del proyecto de medición con el que se asocia la medida (*PRJ_1*), c) El identificador de la categoría de

entidad a la que corresponde la medida (*EC1*), y d) La entidad a la que le pertenece la medida (*Peter*). Es importante mencionar que los identificadores son informados en la definición del proyecto al momento de intercambiarse mediante BriefPD. De este modo, el adaptador de medición genera este mensaje conociendo los vínculos entre métricas/sensores y entidad/contexto. El círculo con un tres en su interior indica la métrica asociada con la medida que viene a implementar, es decir, utilizando la definición del proyecto se sabe qué atributo de la entidad se está informando (En este caso, *dm*_ctemp que indica la temperatura corporal de Peter). El círculo señalado con un cuatro en su interior representa que se trata de una medida estimada (no determinista) e informa el conjunto de (valores, probabilidad) asociado.

De este modo, y como puede apreciarse en la anterior figura, cada mensaje cincamimis incorpora el valor numérico contextualizado por etiquetas que permiten interpretar su significado en contexto. Sin embargo, las etiquetas mediante XML aportan una sobrecarga al intercambio de mensaje que si bien aportan una importante guía semántica es una complicación para dispositivos con recursos limitados. Por ello, se planteó un formato de intercambio de medidas denominado Brief. Este nuevo formato, mantiene la semántica de las medidas para fomentar su interpretación en contexto pero elimina las etiquetas.

El formato Brief corresponde con una nueva y complementaria forma de intercambiar las medidas, manteniendo fuera los datos complementarios. Se dice complementaria, porque el adaptador de medición puede decidir si generar el mensaje siguiendo el esquema CINCAMIMIS en XML, JSON, o alternativamente, si emplea Brief. Por ejemplo, cuando se requiere informar un dato complementario (ejemplo, una imagen y su posicionamiento geográfico), el adaptador de medición podría emplear CINCAMI/MIS mediante XML. Sin embargo, cuando la información a enviar se asocia solo con medidas, el mensaje Brief es una mejor opción por el ahorro que genera en su tamaño, como puede apreciarse en la Figura 25.

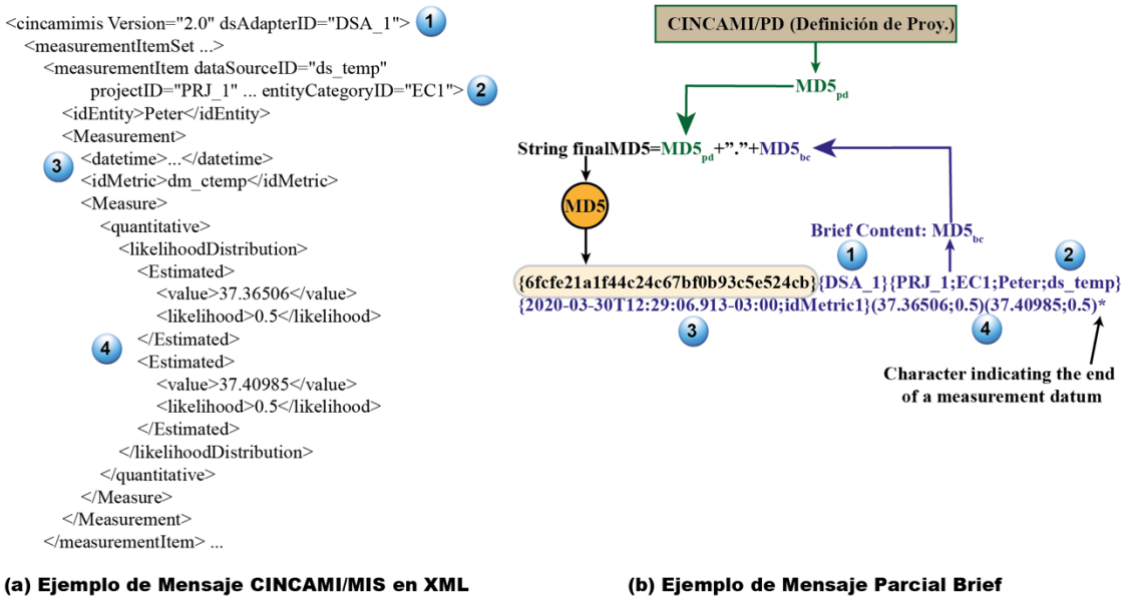

Figura 25 Contraste entre Formatos CINCAMI/MIS en XML y Brief

La anterior figura contrasta el mensaje parcial introducido en la Figura 24 con sus respectiva representación en Brief. El primer punto a resaltar refiere al modo en que el mensaje Brief es generado. Este formato sigue la organización jerárquica de conceptos de acuerdo con el modelo de navegación de CINCAMI/MIS introducido en la Figura 21, Figura 22 y Figura 23 alineado con ECINCAMI. En forma análoga a BriefPD para la definición de proyecto, el hecho de conocer el modelo de navegación y su jerarquía, le permite prescindir de las etiquetas. Los datos que describen alguna información de las medidas se describen entre llaves (“{}”). Se volverá sobre el cálculo de la huella MD5 luego de introducir el contenido del mensaje.

Posicionado en el nivel superior de la jerarquía (Ver círculo con un uno en Figura 25.a), se incorpora el identificador del adaptador de mediciones entre llaves (Ver círculo con un uno en Figura 25.b). Seguido, se desciende en la jerarquía hasta *measurementItem* donde se provee información del proyecto, la categoría de entidad, la entidad, y la fuente de datos (Ver círculo con un dos en Figura 25.a). Esta información se indica en dicho orden, separado por punto y coma y encerrada entre llaves (Ver círculo con un dos en Figura 25.b). Luego, el siguiente paso es describir el momento en que se obtiene la medida junto la métrica asociada (Contraste los círculos con un tres entre la Figura 25.a y Figura 25.b). Dado que las medidas podrían ser estimadas o deterministas, estas se expresan como pares de (valor; probabilidad), indicando el final de la medida (o ser de valores estimados) con un asterisco. Adicionalmente, el asterisco indica que cualquier cadena posterior representa información sobre otra medida siguiendo el mismo orden estricto del modelo de navegación descrito. De este modo, Brief es un formato de orden fijo y con contenido variable. La Tabla 20 describe la estructura y orden seguido por un mensaje Brief, el cual es codificado siguiendo el sistema UTF-8.

Tabla 20 Estructura de la Medida en un Mensaje Brief

Orden	Separador	Descripción
	{	Abre la primera sección describiendo el adaptador de medición empleado para recolectar el dato.
1		Se informa el ID del adaptador de medición como cadena.
	}	Cierra la primera sección indicando el fin del identificador del adaptador de medición.
	{	Abre la segunda sección para describir la información del proyecto de medición.
2		El identificador de proyecto de medición se indica como cadena.
	;	Separa el identificador de proyecto respecto a la categoría de entidad.
3		El identificador de la categoría de entidad se informa como cadena.
	;	Separa el identificador de la categoría de entidad respecto de la entidad.
4		El identificador de la entidad.
	;	Separa el identificador de la entidad respecto del identificador de la fuente de datos (o sensor)
5		El identificador de la fuente de datos (o sensor)
	}	Cierra la segunda sección que describe la información del proyecto.
	{	Abre la sección que describe la métrica e instante de la medida.
6		Una representación tipo cadena para la instancia de ZonedDateTime de la plataforma Java 8.
	;	
7		El identificador de la métrica expresado como cadena
	}	Cierra la sección que describe la métrica y el instante de la medida.
…	(	Indica el inicio de la descripción de un valor.
8…		Indica el valor medido como una instancia de BigDecimal expresado como cadena.
…	;	
9…		Indica la probabilidad asociada con el valor de la medida. Para una medida determinista, la probabilidad es 1. Sin embargo, para un valor estimado, el valor podrá estar dentro del intervalo (0; 1).
…	)	Indica el final de la descripción del valor
	*	Indica el final de la descripción de la medida.

El asterisco al final del mensaje puede representar dos situaciones. Por un lado, cualquier carácter encontrado será interpretado como el inicio de la descripción de una nueva medida siguiendo el mismo orden de descripción. Por otro lado, de no existir ningún carácter, señala el fin del mensaje.

La huella MD5 de CINCAMIMIS se calculaba en base al contenido de la etiqueta *measurementItem* (Ver círculo 2 en la Figura 24). La debilidad de ello es que focaliza en el contenido de la medida pero no sobre la definición de proyecto en la que se sustenta. Por ello, el mensaje Brief incorpora una sutil variación para su cálculo que contempla la definición del proyecto de medición que el adaptador de medición emplea para generar el mensaje (por ejemplo, del cual toma los identificadores). Así, Brief calcula el MD5 a

partir de una cadena que contiene los MD5 desde la definición del proyecto y contenido informado respectivamente separado por un punto (Ver Figura 25.b). El MD5 resultante es incorporado como encabezado del mensaje Brief lo que permite verificar la integridad del mensaje en términos de la medida pero ahora también en función de la definición de proyecto en la que se sustenta.

La librería cincamimis disponible bajo los términos de la licencia Apache 2.0 en GitHub (https://github.com/mjdivan/cincamimis), incorpora una implementación de referencia para CINCAMI/MIS mediante JSON y XML como así también del formato Brief para su contraste. Dicha implementación será retomada en el siguiente capítulo al momento de introducir los árboles de Merkle en la transmisión de datos [146].

5.2.2 Recolección de Datos Distribuida

La recolección de datos en campo se distribuye y organiza alrededor de las figuras del adaptador de mediciones y las pasarelas, tal y como se introdujo en la Figura 20. La Figura 26 detalla la organización interna de pasarelas y adaptadores de medición.

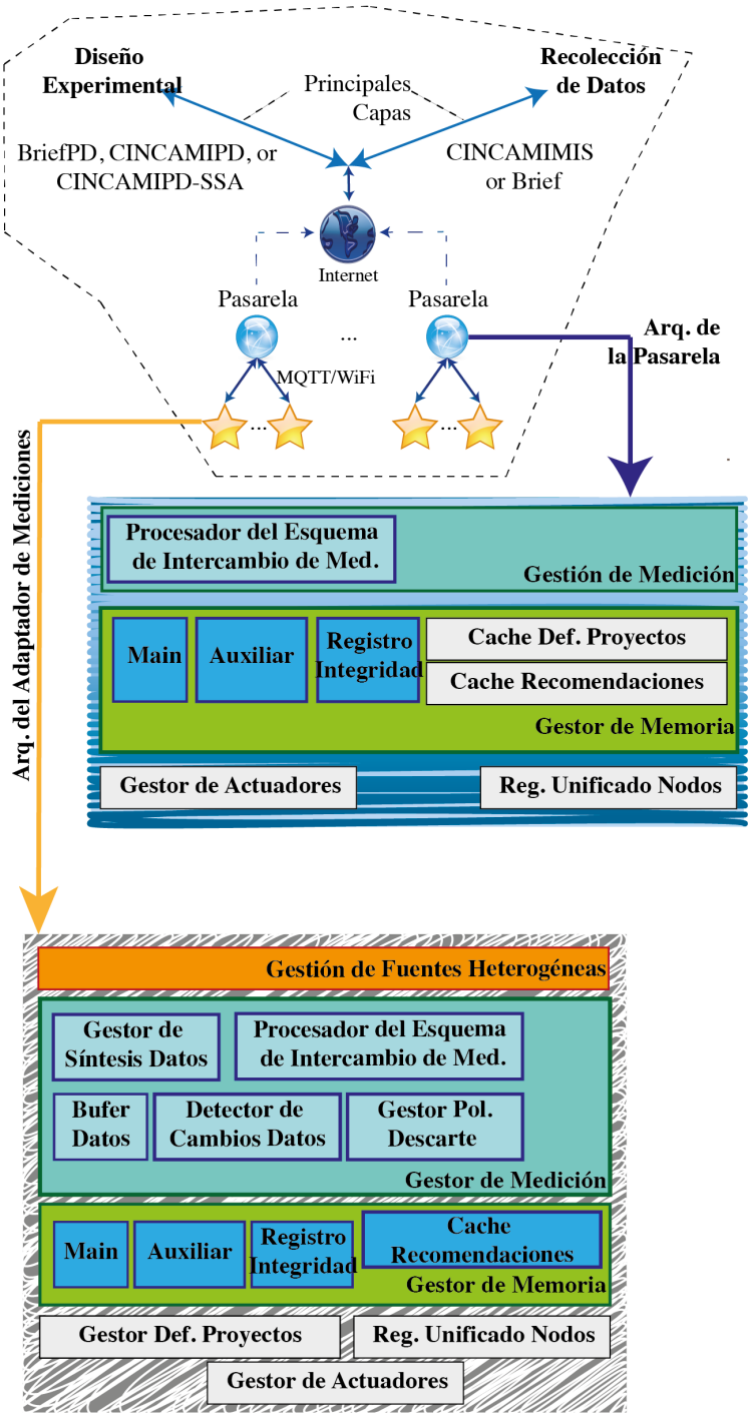

Figura 26 Organización de los Recolectores de Datos

La capa de organización de datos se estructura en dos niveles. El nivel más bajo se asocia con el adaptador de medición (o puente semántico), el cual es responsable de articular sensores de datos heterogéneos con las pasarelas. El nivel superior se vincula con las pasarelas, las cuales actúan como concentradores e intermediarios entre los adaptadores de medición y las capas de la arquitectura en la nube. Las pasarelas

incorporan funcionalidades adicionales orientadas a proveer servicios locales a los adaptadores de medición.

El adaptador de medición se organiza como una topología en estrella debido a su inmediatez respecto de los sensores. Es decir, el centro lo representa el adaptador comúnmente ejecutándose sobre una computadora de placa simple (en inglés, *Single Board Computer -SBC*), mientras que los extremos se asocian con los sensores.

Las pasarelas siguen una organización de árbol multicamino, donde las pasarelas constituyen la raíz del árbol y cada adaptador de medición representan hojas. Debido a la complejidad de las capas de recolección de datos, cada adaptador/pasarela podría ser analizado como un conjunto de subcapas. Como se mencionó, el adaptador es responsable por articular las fuentes de datos heterogéneas con el resto de las capas de la arquitectura. Medidas y metadatos se envían a la capa de recolección de datos en la nube a través de la pasarela para su procesamiento (Ver Sección 5.1). Las subcapas del adaptador pueden describirse como sigue:

- **Gestor de Metadatos**: Existen tres responsabilidades albergadas en este nivel. La primera se asocia con la gestión de las definiciones de proyectos de medición en la cual el dispositivo (y sus sensores conectados) está participando. La segunda se vincula con el conjunto de dispositivos involucrados en la estrategia de recolección, indicando el rol de cada uno de ellos para saber cuáles son activos y los bloqueados (por ejemplo, debido a riesgos de seguridad). La tercera responsabilidad regula el modo en que las acciones informadas serán tomadas a través de los actuadores locales (cuando ellos están disponibles en el adaptador).

- **Gestor de Fuentes Heterogéneas**: Su objetivo reside en emparejar cada sensor conectado al SBC con el identificador de métrica en términos del proyecto de definición. Este aspecto es esencial para emparejar cada dato proveniente de los sensores respecto del concepto que cuantifica.

- **Gestor de Memoria**: Dado que el adaptador de medición (AM) puede actuar colaborativamente con otros adaptadores de medición cercanos, la memoria se desglosa en cuatro.
    1) El componente principal contiene las medidas recolectadas desde los sensores directamente vinculadas al AM exclusivamente.

    2) El componente auxiliar reserva una porción para colaborar con AM cercanos.

    3) El registro de integridad mantiene un seguimiento de las medidas informadas a través un árbol de Merkle [146]. Este aspecto en particular se abordará en el siguiente capítulo.

4) El caché de recomendación contiene un conjunto con las recomendaciones más probables basada en las situaciones descrita para las últimas medidas informadas.

- **Gestor de Medición**: Esta capa se compone de varias responsabilidades, a saber:
  1) El Buffer de Datos estructura y enlaza medidas y metadatos desde los sensores.

  2) El Detector de Cambio de Datos es responsable por detectar fluctuaciones o cambios en la distribución de datos utilizando diferentes políticas.

  3) El Gestor de Política de Descarte de Datos regula las prioridades de retención de datos siguiendo las políticas actuales. Por ejemplo, ante una potencial sobrecarga de datos, deben retenerse las medidas asociadas con un conjunto dado de métricas.

  4) El Gestor de Síntesis de Datos resume datos para evitar transmisiones de datos improductivas. Por ejemplo, cuando una métrica no presenta variación, permite sintetizar e informar un resumen en lugar de los detalles.

  5) El Procesador del Esquema de Intercambio de Mediciones es responsable por articular búfer de datos, los metadatos de medición (dado en la definición del proyecto), y las medidas en memoria para generar el flujo de medidas homogéneo (sea en XML, JSON, o Brief).

Las pasarelas introducen un menor número de capas en la Figura 26 debido a que ellas no poseen contacto directo con los sensores. Ellas están principalmente orientadas a servir como nodo intermedio y soportar a los AM conectados. Su funcionalidad puede ser categorizada en tres capas:

- **Gestor de Metadatos**: Este incorpora el registro unificado de nodos y el gestor de actuadores. Dado que la pasarela relaciona un conjunto de AM, contiene un conjunto de actuadores supervisados, conociendo la relación entre actuador y AM.

- **Gestor de Memoria**: Este se organiza en relación a cinco componentes. Los conceptos subyacentes para la memoria principal, auxiliar, registro de integridad, y caché de recomendaciones son análogos al AM. Sin embargo, la pasarela contiene datos desde todos los AM directamente enlazados.

Adicionalmente, contiene un gestor de proyectos de definición que articula las últimas definiciones disponibles para los AM vinculados. De este modo, cuando una definición de proyecto es actualizada, la pasarela ejecuta una única actualización, haciéndola disponible a los AM.

- **Gestión de Medición:** Se focaliza en procesar datos y metadatos desde los AM para informarlos a la capa de recolección en la nube, regulando y optimizando las transmisiones de datos.

5.2.3 Transmisión Indirecta de Medidas

Un AM puede transmitir medidas en forma directa a la nube, mediante las pasarelas u otro AM. Adicionalmente, la pasarela puede concentrar un conjunto de medidas desde diferentes AM para optimizar la transmisión de datos a la nube. Esta sección introduce detalles para la transmisión indirecta entre AM. Adicionalmente, a partir de los resultados de una simulación discreta, se describen patrones de referencia para las operaciones de envoltura de mensajes CINCAMIMIS [147].

Tanto las pasarelas como AM poseen un Registro Unificado de Nodos (RUN). Los principales campos del registro se sintetizan en la Tabla 21.

Tabla 21 Principales Campos del Registro Unificado de Nodos

Campo	Descripción
MA_ID	Identificador del AM o pasarela.
Footprint	Una huella MD5 calculada a partir de MA_ID, ANC, ADS, GML, y el rol.
ANC	Acrónimo en inglés de *Authorized Network Cards* (Tarjetas de Red Autorizadas).
ADS	Acrónimo en inglés de *Authorized Data Sources* (Fuentes de datos autorizadas).
Role	Este define el comportamiento a seguir por el AM o pasarela.
GML	Acrónimo en inglés para *Geographic Markup Language* (Lenguaje de Marcado Geográfico)

El objetivo de RUN es identificar cada fuente de datos o intermediario extendido en el campo de recolección de datos. Se dice que es unificado dado que el registro en PAbMM es compartido entre todos los proyectos de medición. Esto es así para fomentar el uso compartido de fuentes de datos como se introdujo en la Figura 19.

El campo *footprint* cambiará en caso de que nuevas fuentes de datos o tarjetas de red sean incorporadas. En caso de discordancia con la huella, por ejemplo, otro AM podría rechazar una solicitud de transmisión indirecta de medidas. En cuanto al rol de cada nodo, este puede ser uno de los siguientes:

- **Data Collector**: El nodo solo informa medidas provenientes de sus fuentes de datos (sensores) directamente conectados.
- **Gateway**: El nodo limita su funcionalidad a transmitir medidas en nombre de otros AM. No produce ninguna medida que le sea propia.

- **Cooperative**: El nodo informa sus propias medidas provenientes de sus fuentes de datos (o sensores) y colabora con otros AM para retransmitir sus medidas.
- **Blocked**: El nodo no se encuentra autorizado a interactuar con el procesador de datos (ejemplo, la nube) u otro nodo.

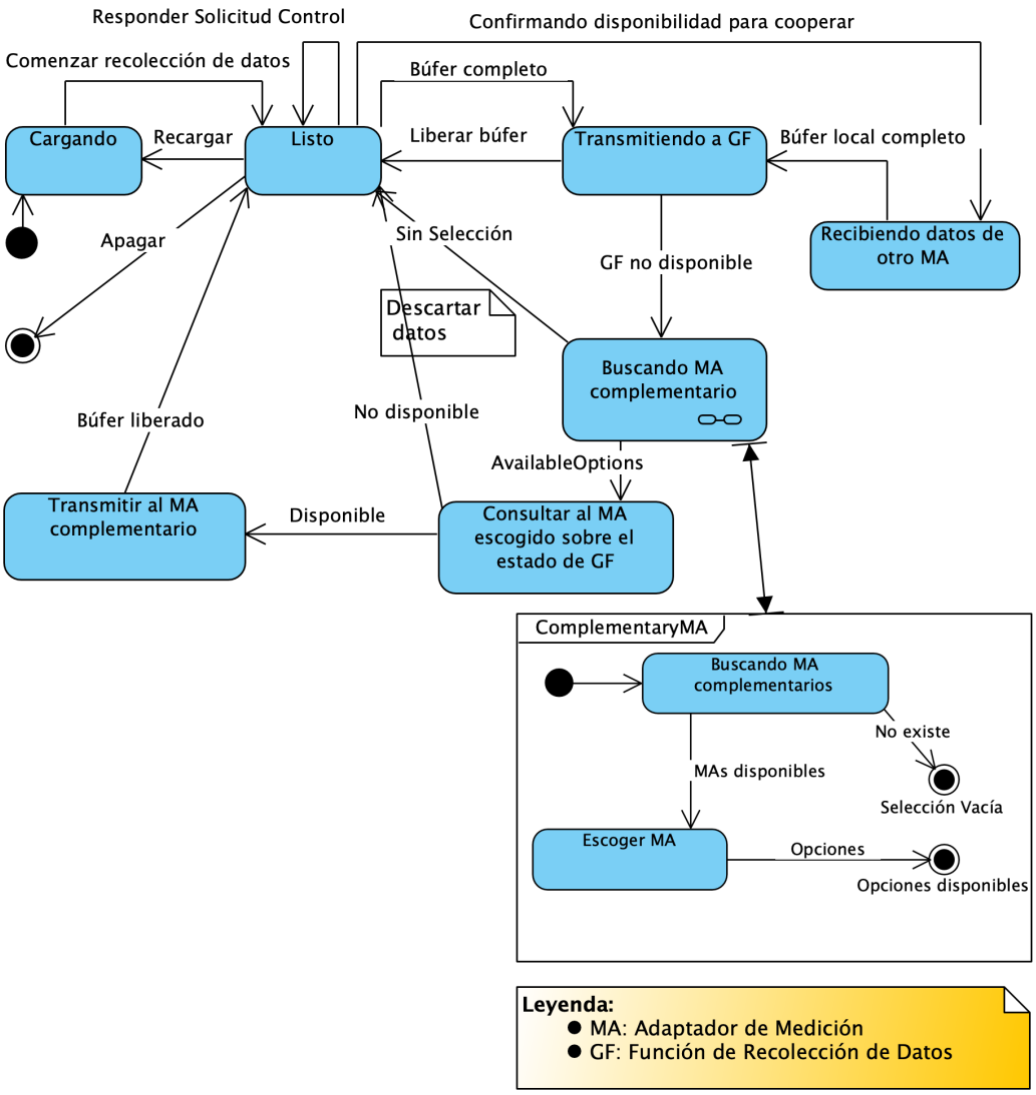

Figura 27 Diagrama de Estados para los Nodos

La Figura 27 describe los diferentes estados por los que puede transitar un nodo en la recolección de datos. El estado "*Cargando*" se asocia con la carga de la definición del proyecto de medición (es decir, el contenido del mensaje BriefPD) para identificar la entidad a monitorear, métricas, fuentes de datos, su emparejamiento, etc. Una vez cargada la información, el AM está listo para recibir medidas desde las fuentes de datos (es decir, sabe cómo articular los valores numéricos con cada metadato para generar el esquema de intercambio de mediciones). Eventualmente, podría recargar la definición de proyecto si agregare uno nuevo (ejemplo, para colaboración) o actualizare la versión

del vigente. Durante el estado "*Listo*", el AM es capaz de responder a solicitudes de control (por ejemplo, una solicitud de cooperación desde otro AM) y recolectar medidas desde los sensores. Al llenarse el búfer local (y de no mediar otra política de transmisión), el AM transita al estado "*Transmitiendo a GF*" donde intentará transmitir el flujo de medidas (empleando CINCAMIMIS o Brief) a la función de recolección en la nube. En caso de que la función de reunión no responda al intento de transmisión, automáticamente se transitará al estado "*Buscando MA complementario*" para localizar un camino alternativo. Por un lado, en caso de no encontrar alternativa, los datos se descartarán y se volverá al estado "*Listo*". Por otro lado, de existir al menos un AM autorizado capaz de colaborar (es decir, disponible y con el rol Gateway o Cooperative), la transmisión seguirá el orden establecido dado en la respuesta de los colaboradores.

Ahora bien, hasta aquí se ha mencionado el RUN y cómo se comporta el AM para redireccionar las medidas en caso de no poder alcanzar a la función de reunión. Sin embargo, nada se ha dicho sobre cómo organizar el flujo de medidas desde un AM a otro para su retransmisión. De este modo, a los efectos de soportar el comportamiento cooperativo, un mecanismo de envoltura se incorpora a la librería de código abierto cincamimis (responsable de generar el flujo de intercambio de medidas en XML, JSON, y Brief) en GitHub. En la envoltura se incorporan las siguientes etiquetas:

- **Origin**: Representa el nodo que actúa como origen de las medidas, es decir, aquel conectado directamente con las fuentes de datos.
- **originTimestamp**: Indica el instante en el que el AM original deriva las medidas al primer AM (cooperativo o pasarela) de la lista.
- **Lifespan**: Tiempo de vida límite expresado en segundos y contabilizados a partir del originTimestamp.
- **Jumps**: Número de saltos (cambios de AM) realizados desde el originTimestamp.
- **knownMA**: Los adaptadores de medición que han sido visitados a través de la secuencia de saltos. Tiene por finalidad evitar los ciclos cerrados.
- **originalMessage**: El mensaje CINCAMI/MIS original que se intenta retransmitir a través de la envoltura.
- **Fingerprint**: Huella MD5 relacionada al mensaje original.

El campo *fingerprint* permite verificar la integridad del mensaje original. En caso de que el mensaje no satisfaga la verificación de integridad, se descarta. De verificar, la etiqueta *jumps* es utilizada para determinar si es posible recibir el mensaje original de acuerdo con la política de datos local del AM (por ejemplo, puede limitar la cantidad máxima de saltos tolerados). Verificados la integridad y saltos, se verificará el atributo lifespan y originTimestamp. La idea es cooperar en la medida que los datos sean útiles y no expirados. De hecho, knownMA es empleado para evitar solicitar dos veces cooperación a un mismo nodo para el mismo mensaje.

Para analizar los tiempos involucrados con el ensobrado y transmisión de los datos, se llevó adelante una simulación discreta para la funcionalidad adicionada en la librería cincamimis sobre un MacBookPro con MacOS Mojave, con 16GB de RAM y procesador Core i7 de 2.9 Ghz. La simulación tuvo por finalidad analizar el tiempo consumido por un MA para:

- Ensobrar o envolver un mensaje variando el número de medidas por mensaje entre 100 y 5000.
- Derivar un mensaje con 500 medidas durante 5 minutos.

El tiempo consumido se compone de los tiempos para calcular la huella MD5, generar la envoltura al mensaje original usando el formato de datos JSON, y comprimir/descomprimir el mensaje envuelto.

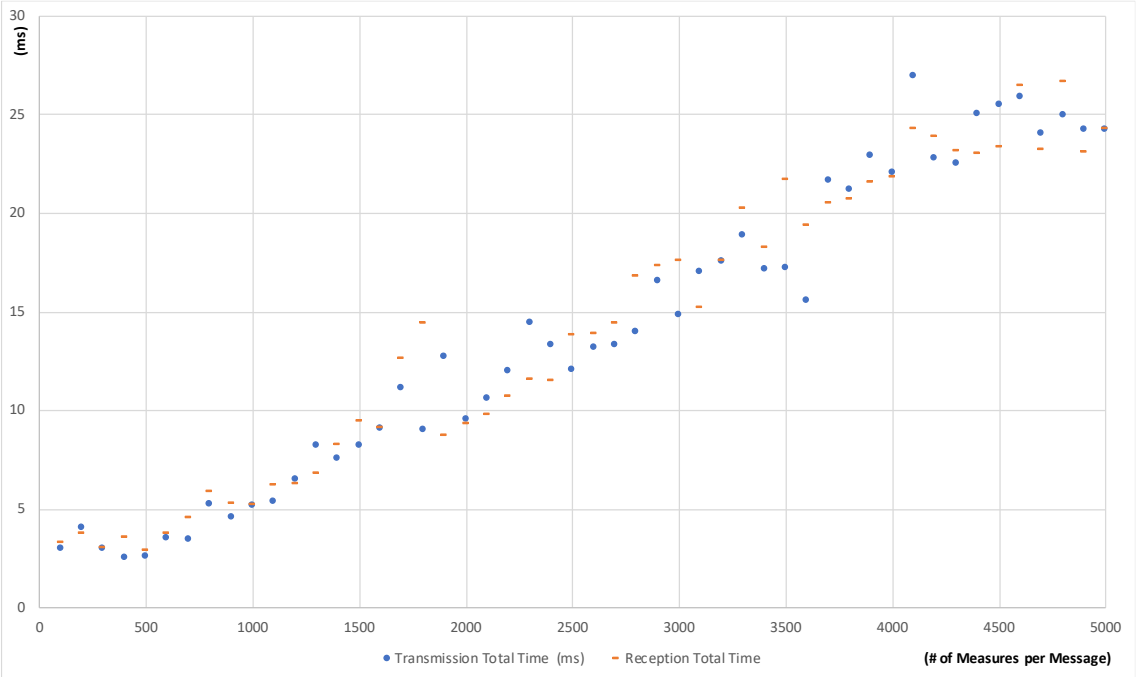

Figura 28 Tiempo de Transmisión vs Recepción para la Estrategia de Envoltura de Mediciones entre 100 y 5000 Medidas por Mensaje

La Figura 28 expone el contraste entre los tiempos de transmisión y recepción (ordenada) del flujo de medidas cuando varía el número de medidas por mensaje (abscisa). Como se puede apreciar, el comportamiento entre transmisión y recepción sería aproximadamente lineal respecto del volumen de medidas a transmitir.

Sin embargo, la Figura 29 describe el comportamiento de los tiempos de transmisión y recepción para un mensaje de 500 medidas durante cinco minutos en forma continua. La media aritmética de los tiempos totales de transmisión fue de 2,75 ms con una desviación estándar de 0,487 ms, una mediana de 2,565 ms y una media recortada (95%) de 2,566. De este modo, contrastando la mediana con su media recortada podría observarse el efeto de los valores atípicos (outliers) en la media aritmética. En otras

palabras, los valores extremos terminan por influir en el cálculo de la media aritmética incrementándola.

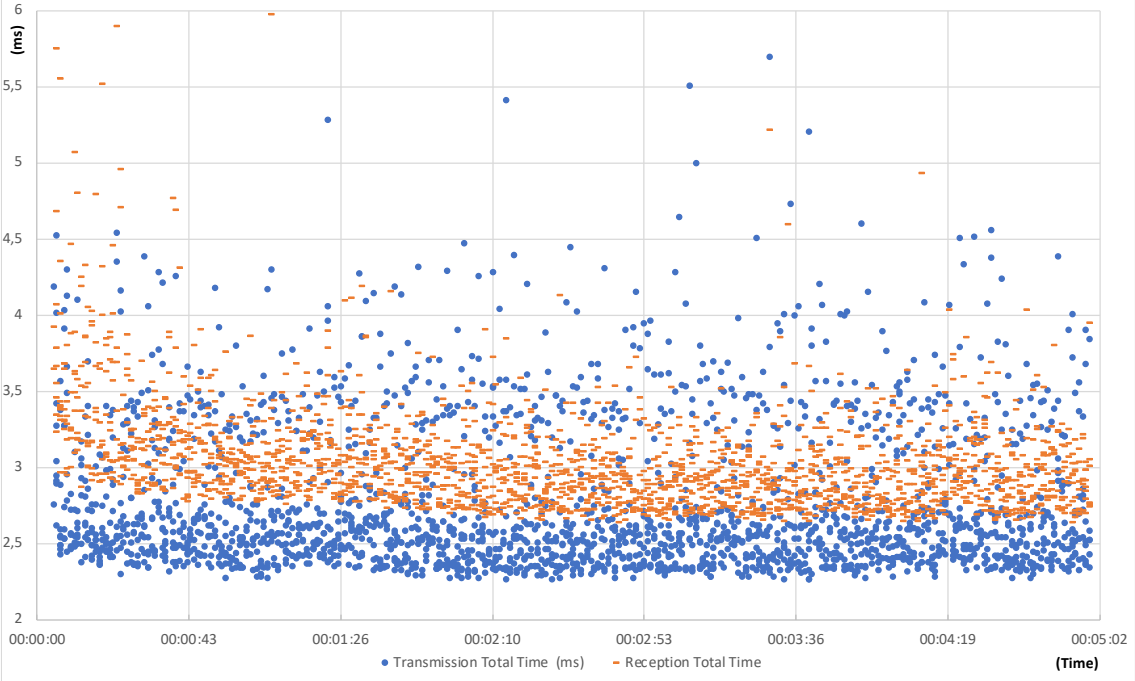

Figura 29 Tiempo de Transmisión y Recepción para 500 medidas por mensaje durante cinco minutos

Algo similar sucede con los tiempos de recepción, cuya media aritmética es 3,031 ms con una desviación estándar de 0,328 ms, una mediana de 2,961 ms y una media recortada (95%) de 2,961 ms. Durante los cinco minutos de la simulación, 2187 mensajes fueron procesados y medidos, lo que implica alrededor de 7,29 mensajes por cada segundo. Ello permitiría indicar que la estrategia de envoltura es factible de ser aplicada sin una sobrecarga significativa. Debe destacarse que se trata de una opción, es decir, un AM puede optar o no por emplear este esquema de transmisión indirecta. Pero de necesitar emplearlo, se estaría incorporando un tiempo total de 2,565 ms para transmitir y 2,961 ms para recibir el mensaje (incluye tiempo de verificación, envoltura, compresión y descompresión).

5.3 Procesamiento de los Estados de Entidad y los Escenarios de Contexto

Hasta el momento se ha introducido desde cómo organizar y estructurar los proyectos de medición, hasta cómo intercambiar mediante el esquema de intercambio de mediciones usando XML, JSON, Brief las medidas con la nube o bien en forma directa por la estrategia de envoltura. Sin embargo, no se ha dado sobre cómo ese flujo de medidas con metadatos se emplea para estimar las probabilidades de estados de entidad y los escenarios al momento en que se procesan.

Esta sección abordar el modo en que el flujo de medidas (datos y metadatos) se procesan para calcular las probabilidades asociadas con ellos y se esquematiza continuando con el caso del paciente ambulatorio introducido en la Figura 17.

La Figura 30 describe una conceptualización para el cálculo de la probabilidad empírica articulando la definición de proyecto dada por BriefPD y el esquema de intercambio de mediciones generada a partir de ella [112].

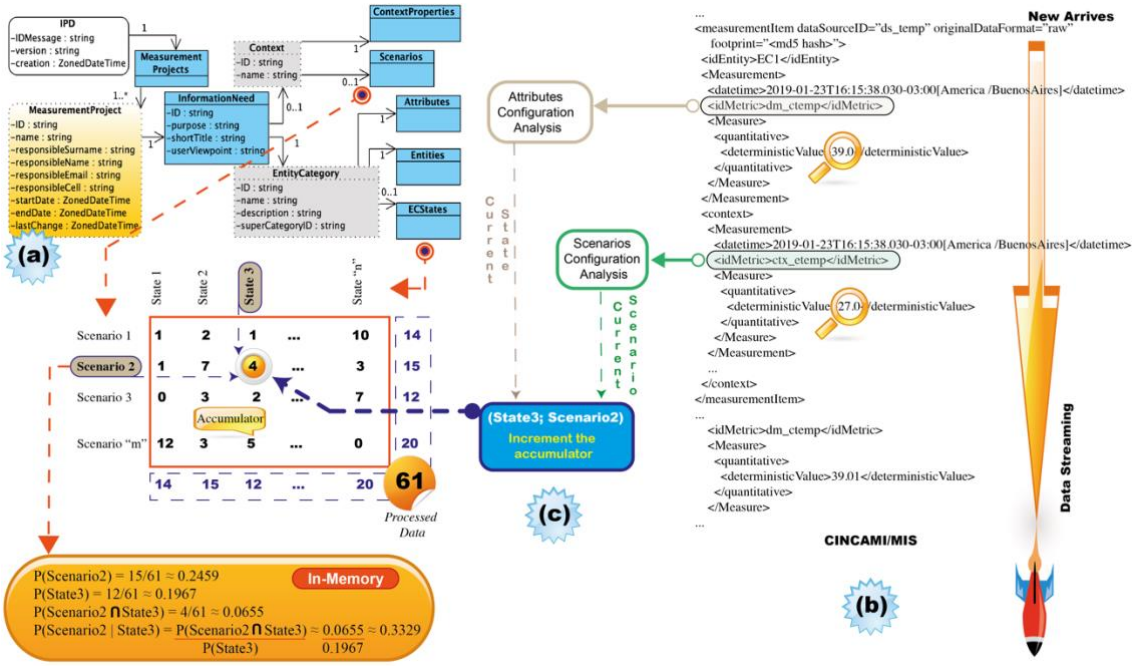

Figura 30 Conceptualización del Cómputo de la Probabilidad Empírica en Tiempo Real

Por un lado, la probabilidad teórica la define el director de proyecto como experto en el área bajo monitoreo. Por otro lado, la probabilidad empírica se calcula en tiempo real considerando los datos que arriban para su procesamiento. De este modo, cuando la definición de proyecto es informada mediante el archivo BriefPD, se crea una matriz bidimensional en memoria con tantas filas como escenarios se definan y tantas columnas como estados se describan. La matriz contiene acumuladores para cada combinación de escenarios y estados de entidad. Cada acumulador se inicializa en cero y representan la ocurrencia conjunta del par (Escenario, Estado) en un proyecto dado.

Una vez creada la matriz e inicializada la memoria para un proyecto dado, El procesador de datos comienza a incrementar los acumuladores basado en el flujo de medición leído desde el esquema de intercambio de mediciones que recibe desde los adaptadores de medición (o pasarelas). Cada proyecto posee su matriz de ocurrencia que se actualiza en forma independiente del resto de los proyectos.

Así, mientras cada flujo arriba se leen las medidas relacionadas con las métricas, se localizan los atributos (o propiedades de contexto) relacionadas para cada uno desde la definición de proyecto en memoria (dada por BriefPD indicado con la estrella dos en la Figura 30). A partir de allí, la combinación de atributos y sus correspondientes medidas

permiten determinar el estado actual de la entidad en términos de la definición dada. Análogamente, las medidas relacionadas con las propiedades de contexto permiten identificar el escenario actual basado en su definición.

Una vez obtenido el estado de entidad y escenario actual, se incrementa el acumulador de la matriz asociado con la combinación entre estado y escenario para representar la ocurrencia conjunta (Ver Figura 30.c). De este modo, la matriz de ocurrencia provee información complementaria como retroalimentación para el modelo de transición de estados (escenarios y estados de entidad). Mientras los modelos de transición mantienen la probabilidad de transición entre el mismo concepto (es decir, entre estados, o bien, entre escenarios), la matriz de ocurrencia provee una perspectiva diferente debido a que provee la ocurrencia conjunta entre escenarios y estados de entidad.

De este modo, la matriz de ocurrencia provee información sobre:

- *La probabilidad empírica actual para un escenario dado*: A partir de la Figura 30, la ocurrencia del "scenario2" para cada estado de entidad es 15 (es decir, 1+7+4+3). Dado que el total de medidas procesadas es 61 para el ejemplo, la probabilidad empírica basada en frecuencia es el cociente entre 15 y 61 (alrededor de 0,2451).
- *La probabilidad empírica actual para un estado de entidad*: En forma análoga al escenario, la ocurrencia del "state3" (Ver su columna asociada en la Figura 30) es 12 (es decir, 1+4+2+5). Dado que el total de las medidas procesadas es 61, la probabilidad empírica para el estado de entidad es el cociente entre 12 y 61 (alrededor de 0,1967).
- *La probabilidad empírica de ocurrencia conjunta entre estado de entidad y escenario*: cada celda de la matriz indica la intersección entre un estado de entidad y otro escenario particular. Su valor numérico es un conteo de la ocurrencia conjunta de ellos. Así, de 61 medidas procesadas en el ejemplo de la Figura 30, 4 corresponden a la ocurrencia conjunta de (Scenario 2, State3). Por tal, la probabilidad conjunta se obtiene como el cociente entre 4 y 61 (es decir, alrededor de 0,0655).
- *La probabilidad condicional del par (Escenario, Estado)*: A sabiendas que el estado actual de la entidad es "State3", la idea sería estimar la probabilidad de que el "Scenario2" suceda. A partir del teorema de Bayes, necesitaríamos a probabilidad conjunta entre "Scenario2" y "State3" (es decir, el 0,0655) junto con la probabilidad de ocurrencia del "State3" (es decir, el 0,1967). A partir de allí, calculando el cociente entre la probabilidad de ocurrencia conjunta y la probabilidad de "State3", es posible obtener la probabilidad condicional (es decir, 0,0655 / 0,1967 ≈ 0,3329).

La operación de actualizar la matriz de ocurrencia no representa una sobrecarga significativa de procesamiento debido a que se limita a incrementar un acumulador en la matriz, donde el acceso se da en forma directa a partir del escenario y estado actual que se asocian directamente con una fila y columna específica respectivamente.

En general, la matriz de ocurrencia consume una pequeña porción de memoria limitado por el número de escenarios y estados de entidad. Es decir, Si "n" es la cantidad de estados de entidad y "m" es la cantidad de escenarios, una matriz densa consumiría "mxn". Es importante mencionar que los mismos son definidos por el director de proyecto y se tiende a representaciones parsimoniosas (es decir, un conjunto mínimo de estados y escenarios lo suficientemente descriptivos que eviten complejidad innecesaria). Por ejemplo, un proyecto con 10 escenarios y 5 estados empleando un tipo de datos *unsigned long* de JAVA (es decir, 8 bytes) consumiría 400 bytes mantener la matriz en memoria, con capacidad de representar 2^{64}-1 dígitos por cada acumulador.

Sin embargo, es importante mencionar que cuando se ha procesado un volumen de medidas cercano al límite máximo (en este ejemplo, un acumulador con 2^{64}-1 dígitos) es posible indicar al menos dos alternativas al procesador de datos:

1. Mantener una copia en memoria del último estado conocido de la matriz de ocurrencia para referencia del procesador de datos, reinicializando los acumuladores de la matriz actual a cero para continuar el procesamiento. En tal sentido, debe mencionarse que esta alternativa duplica los requerimientos de memoria por proyecto (por ejemplo, si el consumo original fuere los 400 bytes, se requeriría ahora 400*2¡ 800 bytes).

2. Descartar la matriz de ocurrencia vieja, reinicializarla a cero e indicar al procesador de datos que es necesario utilizar la matriz actual como una referencia para su cómputo.

Para ejemplificar el mecanismo de cálculo, suponga que se tiene que monitorear un paciente ambulatorio de 40 años basado en su frecuencia cardiaca, mientras que interesa seguir el efecto de la temperatura y humedad ambiental. La métrica del valor de la frecuencia cardíaca refiere a la entidad, mientras que el valor de la temperatura y humedad ambiental refieren al contexto. La Tabla 22 describe una vista parcial acotada al ejemplo para el indicador relacionado con la métrica de frecuencia cardíaca.

Tabla 22 Definición Parcial del Indicador "Nivel de Frecuencia Cardíaca"

Indicador	**Nivel de la Frecuencia Cardiaca**
Dominio	$c \in \mathbb{N} / 0 < c < 200$
Escala	Ratio
Unidad	Bpm (beat per minute)
Métrica	"Valor de la frecuencia cardíaca" utilizando el método del pulso radial

Indicador	Nivel de la Frecuencia Cardiaca
Criterios de Decisión	< 60 bpm → Riesgo [60; 100] bpm → Bajo Riesgo (100; 120] bpm → Normal (120; 150] bpm → Bajo Riesgo > 150 bpm → Riesgo

Para interpretar básicamente la temperatura y humedad ambiental, se definen sus respectivos indicadores en forma análoga a la frecuencia cardíaca como expone la Tabla 23 y la Tabla 24.

Para el nivel de temperatura ambiental, los criterios de decisión se limitan a tres interpretaciones básicas (frío, normal, caliente). Por otro lado, para el nivel de la humedad ambiental se defines tres interpretaciones básicas (Baja, Normal, Alta).

Tabla 23 Definición Parcial del Indicador "Nivel de Temperatura Ambiental"

Indicador	Nivel de la Temperatura Ambiental
Dominio	$t \in \mathbb{R} \,/\, -10 < t < 50$
Escala	Intervalo
Unidad	ºC (Grados Celsius)
Metrica	"Valor de la Temperatura Ambiental" utilizando el método de flujo de calor
Criterios de Decisión	< 24 ºC → Frío [24; 30] ºC → Normal > 30 ºC → Caliente

Tabla 24 Definición Parcial del Indicador "Nivel de Humedad Ambiental"

Indicador	Nivel de Humedad Ambiental
Dominio	$h \in \mathbb{R} \,/\, 0 < h < 100$
Escala	Intervalo
Unidad	%
Métrica	"Valor de la Humedad Ambiental" utilizando el método de Espejo enfriado
Criterios de Decisión	< 50 % → Baja [50; 60] % → Normal > 60 % → Alta

De este modo, es posible simplificar la definición de escenarios como expone la Tabla 25 a partir de la combinación de interpretaciones de los indicadores de temperatura y humedad.

Tabla 25 Definición Básica de Escenarios

Temperatura/Humedad	**Alta**	**Normal**	**Baja**
Caliente	Peligroso	Peligroso	Precaución
Normal	Precaución	Amigable	Amigable
Frio	Peligroso	Precaución	Amigable

Como se puede observar de la tabla anterior, una vez determinada la interpretación de los indicadores de temperatura y humedad a través de sus criterios de decisión, el escenario asociado queda determinada por la intersección de sus interpretaciones. Es decir, si la humedad es "Alta" y la temperatura es "Caliente" el escenario será

"Peligroso" para esta definición. Ahora bien, el escenario actual y estado de entidad afectan la interpretación de la frecuencia cardiaca de la entidad como expone la Tabla 26. Para el estado de entidad, se ha simplificado a tres estados (En Actividad, Normal, Descansando) basado en la frecuencia cardíaca.

Tabla 26 Definición parcial del Indicador de Nivel de Frecuencia basado en Escenarios y Estados de Entidad

Estados/ Escenarios:	Amigable	Precaución	Peligroso
En Actividad (153; 180]	< 120 bpm → Riesgo [120; 153] bpm → Bajo Riesgo (153; 180] bpm → Normal (180; 190] bpm → Bajo Riesgo > 190 bpm → Riesgo	< 120 bpm → Riesgo [120; 153] bpm → Bajo Riesgo (153; 175] bpm → Normal (175; 185] bpm → Bajo Riesgo > 185 bpm → Riesgo	< 130 bpm → Riesgo [130; 153] bpm → Bajo Riesgo (153; 170] bpm → Normal (170; 180] bpm → Bajo Riesgo > 180 bpm → Riesgo
Normal [90; 153]	< 80 bpm → Riesgo [80; 90] bpm → Bajo Riesgo (90; 153] bpm → Normal (153; 160] bpm → Bajo R. > 160 bpm → Riesgo	< 80 bpm → Riesgo [80; 90] bpm → Bajo Riesgo (90; 145] bpm → Normal (145; 153] bpm → Bajo Riesgo > 153 bpm → Riesgo	< 80 bpm → Riesgo [80; 90] bpm → Bajo Riesgo (90; 135] bpm → Normal (135; 153] bpm → Bajo Riesgo > 153 bpm → Riesgo
Descansando [60; 89]	< 55 bpm → Riesgo [55; 60] bpm → Bajo Riesgo (60; 89] bpm → Normal (89; 95] bpm → Bajo Riesgo > 95 bpm → Riesgo	< 55 bpm → Riesgo [55; 60] bpm → Bajo Riesgo (60; 80] bpm → Normal (80; 90] bpm → Bajo Riesgo > 90 bpm → Riesgo	< 55 bpm → Riesgo [55; 60] bpm → Bajo Riesgo (60; 75] bpm → Normal (75; 85] bpm → Bajo Riesgo > 85 bpm → Riesgo

La Figura 31 describe la definición teórica de probabilidades dadas en el proyecto de medición basado en el modelo de transición de estados y escenarios.

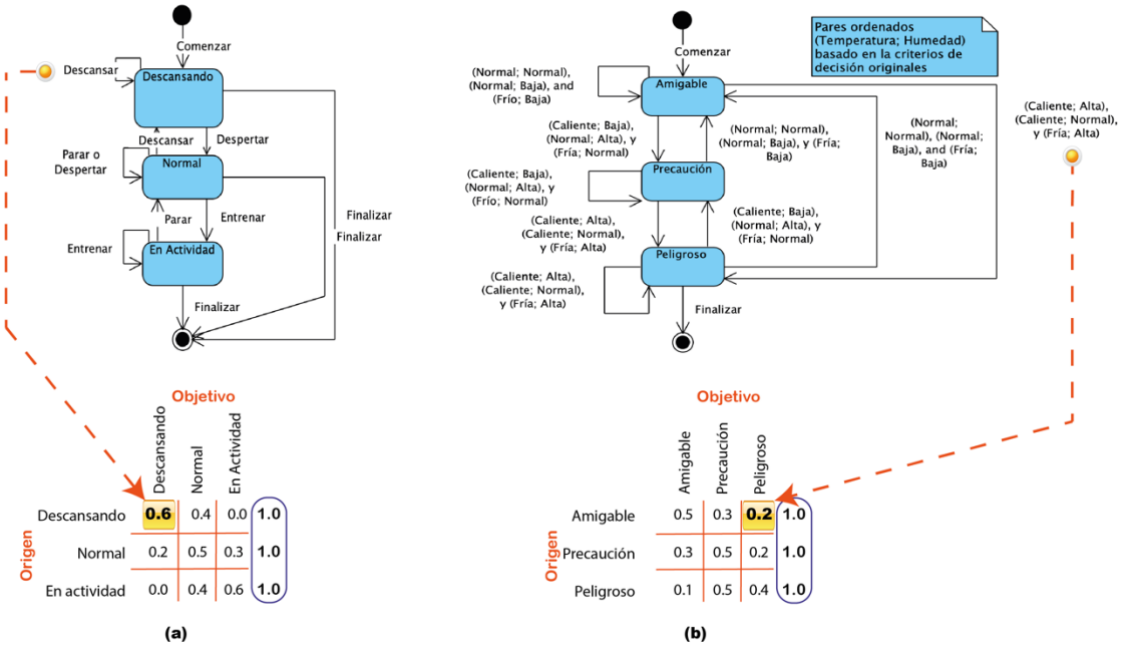

Figura 31 Modelo de Transición y Probabilidades Asociadas para (a) Estados de Entidad y (b) Escenarios

Por ejemplo, el 0.6 indicado en la matriz de la Figura 31.a representa la probabilidad teórica de mantenerse en el estado “Descansando”. Las transiciones en los diagramas refieren a un conjunto limitado de acciones que la entidad puede realizar limitado al ejemplo (es decir, entrenar, parar, o despertar). Sin embargo, los mensajes de la Figura 31.b son sutilmente diferentes porque refieren a pares ordenados basados en un interpretación dada para el par (temperatura, humedad). Por ejemplo, dado un escenario ”Amigable”, la probabilidad para transitar al escenario “Peligroso” es de 0.2, lo cual podría suceder debido a una alta temperatura (es decir (Caliente; Alta) o (Caliente; Normal)), o incluso, una baja temperatura pero con alta humedad (es decir, (Fría; Alta)). Como se puede apreciar en la Figura 31, las probabilidades mencionadas en las matrices refieren solo a transiciones de estados, o bien, solo transiciones de escenarios pero no ambas simultáneamente como introdujo la Figura 30.

Por ello, la Figura 32 introduce un esquema conceptualizando las instancias de procesamiento para implementar el cómputo de las probabilidades empíricas individuales (escenarios y estados), pero también las conjuntas y condicionales.

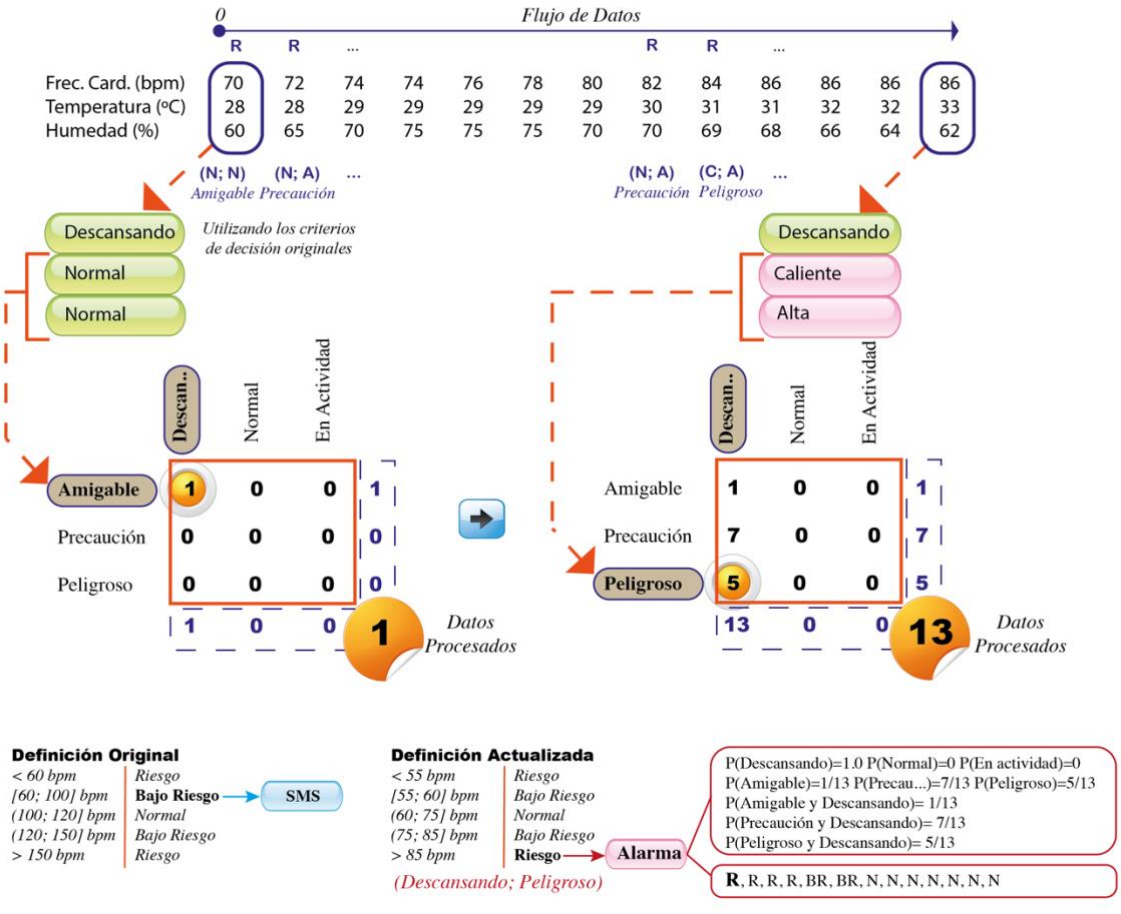

Figura 32 Secuencia Ilustrativa del Comportamiento Esperado de la Arquitectura de Procesamiento

En el ejemplo se cuenta con una matriz 3x3 dado por los escenarios (Amigable, Precaución, Peligroso) y los estados de entidad (Descansando, Normal, En Actividad).

Cada vez que una combinación de medidas se recibe, analiza y actualiza el acumulador de la matriz tal y como describe la matriz izquierda de la Figura 32. Suponga que las ternas en el flujo de datos representa el arribo de nuevas medidas. Así, para la primera terna, la entidad se encuentra “Descansando” (según la frec. Cardíaca) mientras que su escenario es “amigable” dado que temperatura y humedad son normales. De este modo, el acumulador del par (Amigable, Descansando) se incrementa en uno. Este esquema se repite para cada terna en forma sucesiva.

De este modo, los modelos de transición pueden ser ajustados basados en la matriz de ocurrencia. Es decir, la probabilidad teórica de mantenerse en el estado “descansando” era de 0,6 (Ver Figura 31.a), sin embargo, luego de procesar todos los datos del flujo en el ejemplo, no ocurrieron cambios de estados. Por tal, la probabilidad empírica de mantenerse en el estado “Descansando” es 1 (13/13). Análogamente, la probabilidad asociada con los saltos de escenarios de la Figura 31.b podría ser actualizado empleando los datos procesado de la Figura 32. Es decir, se dieron 12 saltos de escenarios como sigue:

- Amigable → Precaución: 1
- Precaución → Precaución: 6
- Precaución → Peligroso: 1
- Peligroso → Peligroso: 4

Esta posibilidad le permite a la arquitectura de procesamiento identificar los estados y escenarios al momento de procesar los datos, al tiempo que puede ajustar las probabilidades teóricas a través del conteo, estimando las probabilidades empíricas respectivas. Adicionalmente, los indicadores puede seleccionar los criterios de decisión que mejor ajusta al escenario y estado actual para poder interpretar en consecuencia.

5.4 Reunión de los Flujos de Medidas

La arquitectura recibe los flujos de datos y metadatos organizados de acuerdo con la ontología ECINCAMI, siguiendo los lineamientos del esquema de intercambio de mediciones. Este puede informarse desde las pasarelas o adaptadores de medición utilizando JSON, XML o el formato Brief. Éste último si bien optimiza el tamaño de transmisión no contempla datos complementarios. La sección anterior adelantó cómo se infieren estados y escenarios a partir del flujo de medidas para actualizar la matriz de ocurrencia y estimar probabilidades empíricas. Esta sección describe sintéticamente el modo en que la reunión de flujos de datos se da entre múltiples proyectos simultáneos.

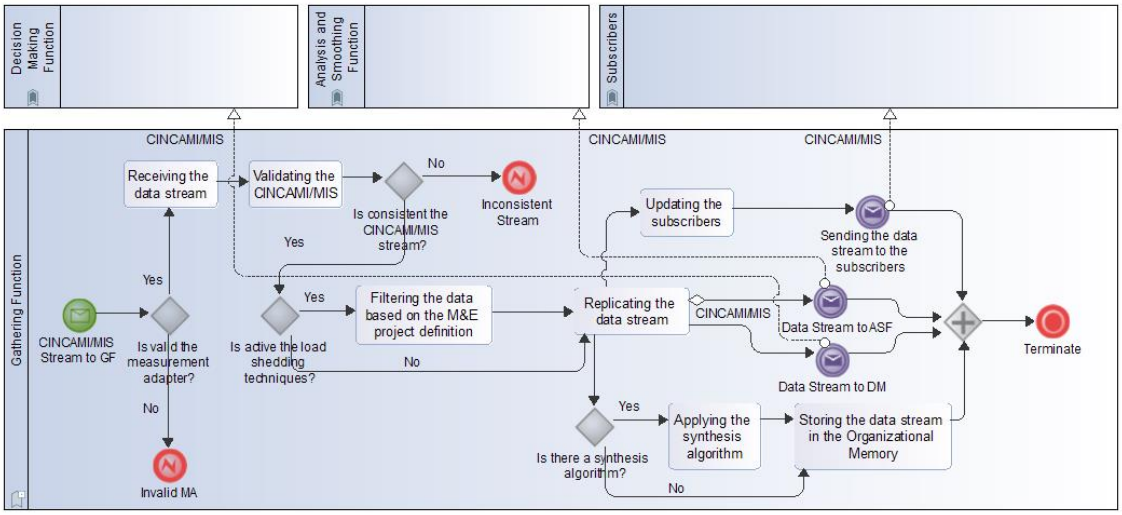

Figura 33 Diagrama BPMN Describiendo la Funcionalidad Esencial de la Recepción de Medidas

La Figura 33 describe la funcionalidad esencial asociada con la recolección general de medidas desde los proyectos de medición. Cuando los flujos de medidas se reciben, estos vienen organizados de acuerdo con el esquema de intercambio de mediciones. La primera instancia de procesamiento evalúa la consistencia del contenido en relación con el adaptador de medición o pasarela. Es decir, se verifica el rol y situación de este para chequear que no se encuentre bloqueado. Si el nodo estuviere bloqueado, el mensaje se descarta sin mayor análisis y el proceso se deriva al estado final señalando “Invalid MA”.

No obstante, cuando el mensaje recibido desde el nodo corresponde con uno no bloqueado (es decir, puede ser actuando bajo el rol cooperativo, pasarela, o recolector de datos) éste se deriva para la verificación de consistencia. Si la versión del mensaje corresponde con CINCAMIMIS expresado como JSON o XML, la verificación corresponde con el cálculo de la huella MD5 basado en el contenido del mensaje. Sin embargo, cuando el mensaje se organiza bajo el formato Brief, la verificación calcula el MD5 para la definición de proyecto y su contenido tal y como se introdujo en la sección 5.2.1. En caso de que la integridad no sea verificada, el proceso culmina en el estado “Inconsistent Stream”.

Los flujos que satisfacen la verificación de integridad se incorporan al registro de integridad global basado en un árbol de Merkle para guardar registro de los flujos recibidos. En paralelo, Se comienza el procesamiento de estados de entidad y escenarios tal como se describió en la sección 5.3. Por otro lado, Se lleva adelante el cómputo de agregaciones y sinopsis para almacenar instantáneas. Sin embargo, puede suceder que la arquitectura tenga su capacidad de procesamiento cercana a su límite máximo, en cuyo caso pueden activarse los mecanismos de descarte de datos selectivos (o en inglés, load-shedding techniques). Estas técnicas permiten interpretar el flujo de medidas de acuerdo con la métrica con las que se asocian y retener aquellas que son prioritarias para un proyecto dado. Esta funcionalidad es selectiva y se activa para mantener el

servicio activo ante situaciones de stress. Es decir, se prefiere la degrades del servicio de recolección ante la indisponibilidad total del mismo.

En caso de no requerirse descarte selectivo, se deriva el flujo a la capa analítica para su procesamiento estadístico mediante lo que se conoce como funcionalidad de análisis y suavizado (en inglés, Analysis and Smoothing Function -ASF) y para la toma de decisiones mediante los clasificadores incrementales basados en árboles de Hoeffding. Adicionalmente, se deriva el flujo para atender a los suscriptores que lo consumen en forma directa mediante el servicio de datos crudos (Ver Figura 20).

De este modo, una vez que se ha procesado y derivado el flujo de medidas por los respectivos canales, este proceso implementado como microservicio llega a su fin.

Como se puede apreciar, el punto neurálgico de esta etapa consiste en el procesamiento del flujo de mediciones desde múltiples adaptadores de medición y pasarelas. La sección 5.2.1 introdujo Brief como alternativa a XML y JSON para el intercambio de mediciones mediante CINCAMIMIS. Se realizaron dos simulaciones discretas:

1. La primera se focalizó en a) Medir el tiempo de generación de mensaje (Brief, XML, y JSON) junto con sus longitudes y tamaños, b) Estimar tiempos y tamaños relacionados con las operaciones de compresión y descompresión de mensaje, c) Tiempo consumido en la traducción del flujo de mensaje al modelo de objetos basado en ECINCAMI. Estas operaciones se realizan para un tamaño variable de mensaje definido arbitrariamente al inicio de la simulación,

2. La segunda simulación mide los mismos aspectos que la primera pero en forma continua durante 20 minutos consecutivos con un tamaño de medidas por mensajes fija (por ejemplo, 200).

La Tabla 27 expone los resultados de tamaños y tasa de compresión de la simulación respecto a un mensaje con 1000 medidas. En la misma, puede observarse la optimización de Brief respecto de los tamaños de mensaje (sean comprimidos o no).

Tabla 27 Tamaños Comparativos para un Mensaje con 1000 medidas

Formato de Datos	Tamaño Normal (KB)	Tamaño Comprimido (KB)	Tasa de Compresión
Brief	642.61	17.50	36.72
JSON	1474.58	42.40	24.77
XML	4000.32	54.20	73.80

La Figura 34 describe diferentes perspectivas para los tiempos de generación del mensaje.

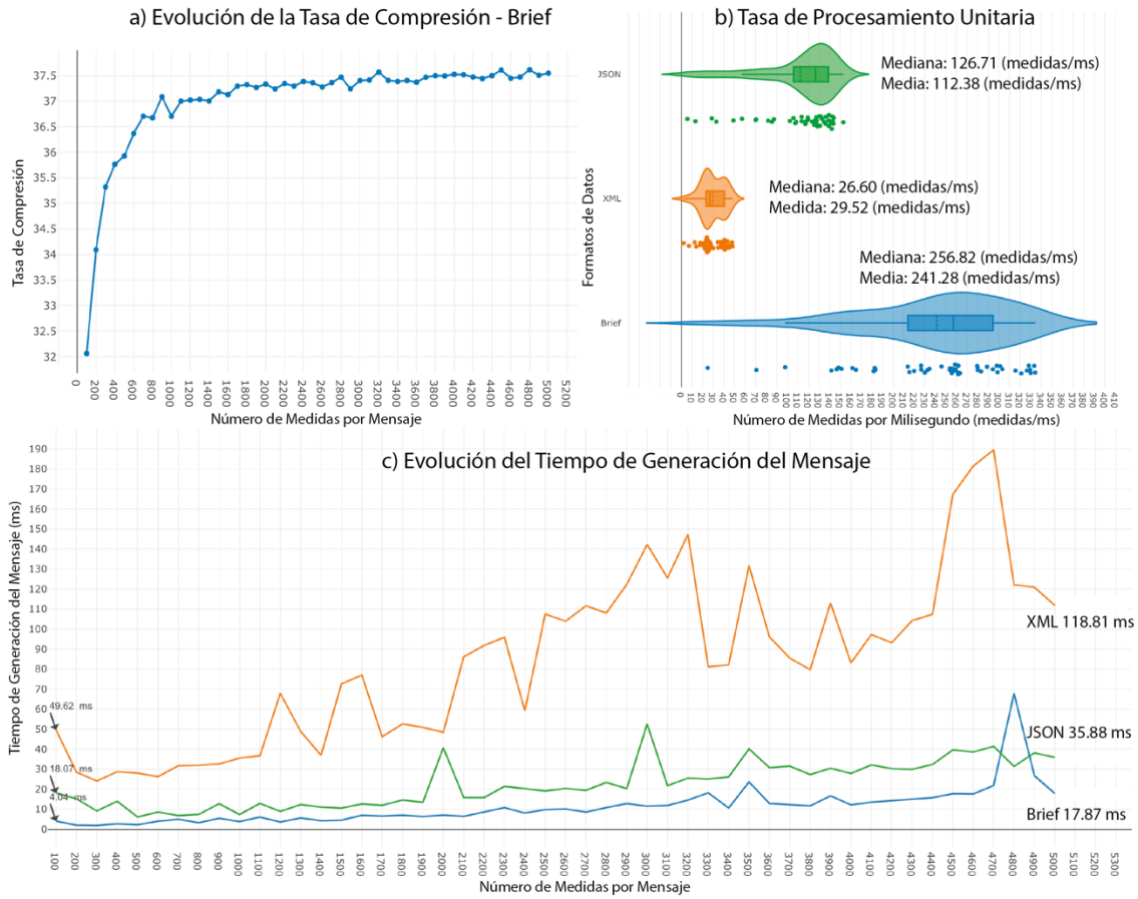

Figura 34 Perspectivas del Tiempo de Generación del Mensaje de acuerdo con el Formato de Datos Empleado

La Figura 34.a muestra la evolución de la tasa de compresión al tiempo que la cantidad de medidas por mensaje crecen, convergiendo alrededor de 37.5. Por otro lado, la Figura 34.b describe la cantidad de medidas por milisegundo que pueden procesarse utilizando cada formato. Allí, la gráfica de violín mostraría una clara superioridad de Brief con una mediana 256.82 medidas por segundo frente a un JSON y XML con 126.71 y 26.60 respectivamente. En otras palabras, el empleo de Brief para la recolección de medidas permitiría procesar el doble de medidas que JSON y casi 10 veces más que XML por cada milisegundo. Una medida de consideración cuando no deben emplearse datos complementarios.

Desde la perspectiva del adaptador de medición, el tiempo de generación del mensaje cobra especial interés además de su tamaño. La Figura 34.c muestra su evolución para distintas configuraciones de tamaño de mensaje. Brief consumiría 17.87 ms para generar un mensaje de 5000 medidas, mientras que JSON y XML consumirían 35.88 ms y 118.81 ms respectivamente. Es decir, Además de las ventajas en tamaño, se deriva que Brief consumiría la mitad del tiempo que JSON y casi siete veces menos que XML.

5.5 Análisis de Datos

La Figura 35 introduce un diagrama BPMN que describe las principales funcionalidades involucradas en la capa analítica (Ver Figura 20) respecto del análisis de los datos.

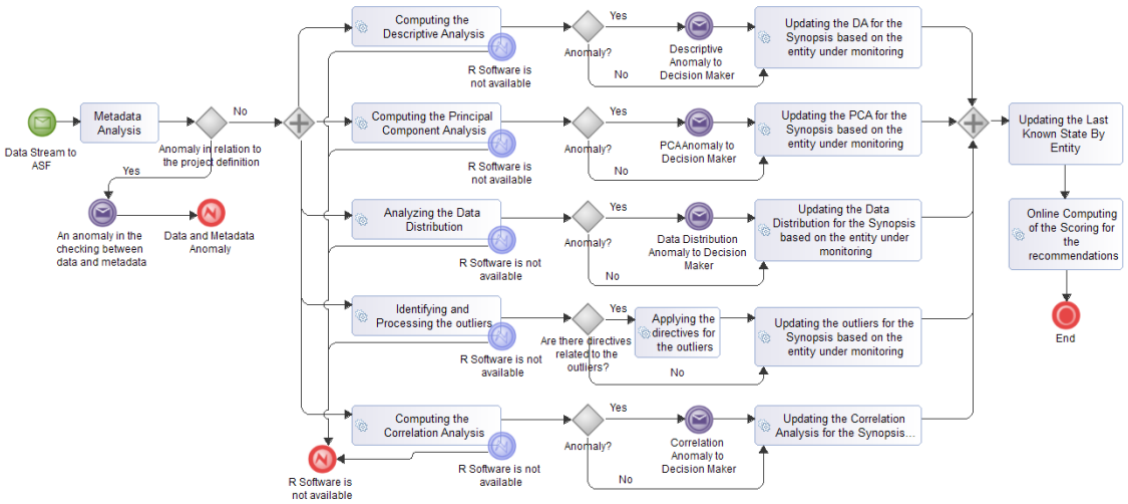

Figura 35 Diagrama BPMN Describiendo los Principales Análisis sobre los Datos

Este proceso analiza el análisis basado en la definición del proyecto de medición y mantiene actualizada síntesis estadísticas y sinopsis en memoria. En este punto, la definición de proyecto de medición le permite conocer a la arquitectura el rango de valores a esperar por métrica. Por ejemplo, la frecuencia cardíaca del paciente no podría ser nunca un valor tal como "-1" ya que sería irracional a su definición.

El proceso lleva adelante cinco análisis estadísticos 1) Análisis descriptivo incremental (por ejemplo, por cada métrica mantiene actualizada la media aritmética, estima la mediana, etc.); 2) Análisis de Componentes Principales (donde todas las métricas - propiedades de contexto y atributos de entidad- corresponden con dimensiones del análisis para estudiar la variabilidad del sistema; 3) Análisis de la Distribución de datos (introducido anteriormente en el análisis de la distancia compuesta); 4) Análisis de Valores Atípicos (univariado y multivariado); 5) Análisis de Correlación entre las métricas involucradas. Estos análisis se desarrollan a nivel de cada proyecto monitoreado por PAbMM.

El resultado de los análisis se actualiza en memoria respecto del último estado conocido del proyecto y sirve de referencia en caso de que el flujo de medidas sea interrumpido. A su vez, la distancia compuesta se actualiza (siguiendo las fórmulas descritas en el capítulo 4) para mantener el scoring (o puntuación) de recomendaciones potenciales dentro y entre proyectos.

5.6 Toma de Decisión y Recomendaciones

La Figura 36 describe un diagrama BPMN articulando los análisis efectuados a partir del flujo de datos como así también el rol de la similitud estructural y comportamental al momento de buscar recomendaciones en caso de detectarse alguna situación tipificada. Para evitar redundancia respecto del capítulo 4 en el que se detalló la estrategia de cálculo de la distancia compuesta, el siguiente diagrama refiere al filtrado del espacio de búsqueda guiado por los coeficientes estructurales y comportamentales.

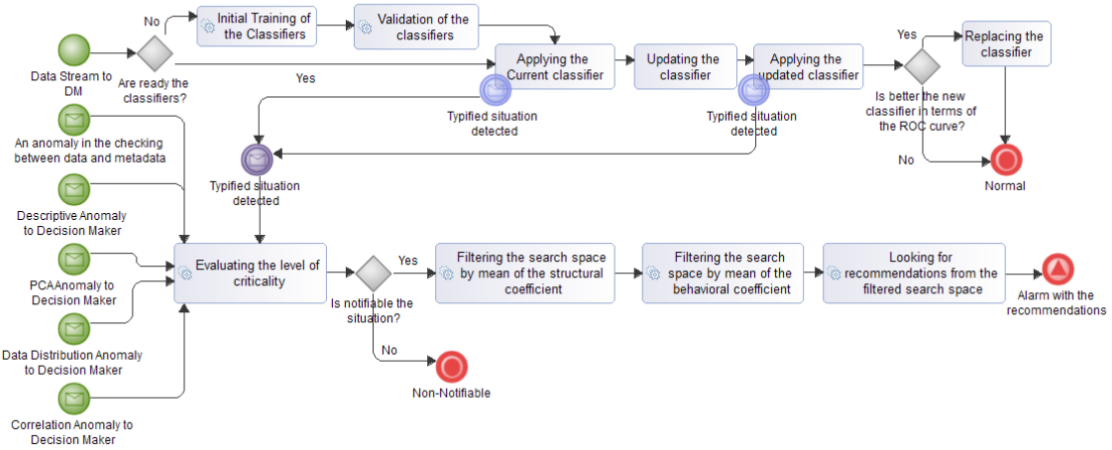

Figura 36 Diagrama BPMN Sintetizando la instancia de Toma de Decisión

Los múltiples puntos de inicios del diagrama refieren a diferentes situaciones que podrían disparar la necesidad de tomar una decisión, y eventualmente, proveer recomendaciones. Por ejemplo, puede deberse a valores atípicos detectados en la serie de datos, comportamiento no esperado para la serie de datos, variabilidad atípica en el sistema, entre otros. En este esquema, los clasificadores son inicializados a partir de un conjunto de datos base empleado para entrenamiento. Luego de lo cual, cada nuevo dato permite actualizar incrementalmente el clasificador basado en un árbol de Hoeffding. El árbol contiene un conjunto de situaciones tipificadas que requieren acción desde el proyecto y se asocia a la combinación de estados y entidades introducidos en la sección 5.3. Por ejemplo, la combinación de un estado de entidad y un escenario pueden corresponder con una “clase” o “categoría” a clasificar por el árbol que requiere una acción (por ejemplo, como se introdujo en la Figura 32, se dispara una alarma cuando la frecuencia cardíaca del paciente excede los 85 bpm, la entidad se encuentra descansando, pero su escenario es “Peligroso” -según la temperatura y humedad ambiental monitoreada-).

Luego de guiar la búsqueda de recomendaciones a partir de los disponibles localmente al proyecto, o bien, de otros mediante la lista ordenada obtenida a través de la distancia compuesta, se informan las alarmas y recomendaciones como etapa final del proceso. Las recomendaciones se organizan en una memoria organizacional

disponible a partir de experiencias previas y conocimiento que se incorpora específicamente para cada proyecto.

5.7 Conclusiones Generales del Capítulo

El capítulo introdujo una perspectiva general de la arquitectura de procesamiento de flujos de datos basado en metadatos de mediciones, articulando la misma con el rol de la definición de proyecto y la ontología ECINCAMI introducida en el capítulo 3.

Se introdujeron diversas mejoras a PAbMM en este capítulo relacionadas con el esquema de intercambio de mediciones, la recolección de datos distribuida, la transmisión indirecta de medidas a través del mecanismo de envoltura, el cómputo de probabilidades empíricas mediante la matriz de ocurrencia conjunta embebiendo el rol de estados de entidad y escenarios junto con la articulación con la distancia compuesta para guiar la reutilización de conocimiento.

En cuanto al intercambio de mediciones, se introdujo el formato Brief que fomenta el intercambio de medidas sin emplear etiquetas (similar a BriefPD para la definición de proyectos) basado en la organización de conceptos derivados de ECINCAMI. Si bien se limita al intercambio de medidas exclusivamente (sin contemplar datos complementarios), se focaliza en proveer una herramienta de intercambio ágil cuando existen limitaciones de ancho de banda y vida de las baterías. A diferencia del mensaje CINCAMIMIS intercambiado mediante XML o JSON, la huella de integridad de Brief contempla tanto el contenido como la definición de proyecto, lo que permite verificar si un conjunto de medidas dadas corresponde o no a una versión dada del proyecto. De este modo, Brief constituye formato complementario a CINCAMIMIS mediante JSON o XML para el intercambio de mediciones. Es decir, cada adaptador de medición o pasarela puede escoger cómo intercambiar el flujo de medidas, la arquitectura los interpretará en consecuencia. Se provee una implementación de referencia de Brief dentro de la librería *cincamimis* disponible en GitHub bajo los términos de la licencia Apache 2.0.

Se introdujo una jerarquía en la recolección de datos distribuida articulando las pasarelas y los adaptadores de medición con roles definidos. Mientras el adaptador de medición es el responsable de establecer el puente semántico entre las medidas y la definición de proyecto para informar a la funcionalidad de recolección datos (nube), las pasarelas se incorporan como un mecanismo de escalabilidad y soporte a adaptadores de medición dispersos en el campo que por sus configuraciones (ejemplo, hardware) no tendrían suficiente alcance o consumirían demasiada batería para actualizar sus definiciones. La organización jerárquica de pasarelas y adaptadores permite contar con diferentes hardware y configuraciones para extender la cobertura en el campo de medición con un balance entre coste y beneficio. De la simulación discreta asociada, se desprende que se logró una mediana de 2,565 ms para el tiempo total de transmisión (generando un mensaje de 500 medidas durante cinco minutos en forma continua) y una mediana de 2,961 ms para el tiempo total de recepción. De este modo, pudo

observarse que una adaptador o pasarela puede colaborar con otros vecinos sin que ello impacte en forma significativa en su procesamiento.

Se implementó un mecanismo de envoltura para el intercambio de medidas a través de otros adaptadores de mediciones y/o pasarelas cuando el origen primario no podía alcanzar (por algún motivo) la recolección de datos en la nube.

Este capítulo introdujo y ejemplificó el modo en que se computan y actualizan las probabilidades empíricas para los estados de entidad y escenarios. Adicionalmente, se introdujo el modo en que las ocurrencias conjuntas entre estados y escenarios se estiman mediante la matriz de ocurrencia, aspecto que originalmente no se encuentran en el modelo de transición de estados (o escenarios) definido por el director de proyecto.

En la capa de recolección de datos de la arquitectura, se describió el impacto que Brief representaba para la tasa de procesamiento de medidas mediante una simulación discreta. Desde la perspectiva del tamaño del mensaje, se pudo observar que Brief consume solo 642,61 KB (o 17,50 KB comprimido) para un mensaje con 1000 medidas versus 1474,58 KB (42,40 KB comprimido) de JSON y 4000,32 KB (54,20 KB comprimido) de XML. Desde la perspectiva de la recolección y procesamiento de medidas, Brief arrojó una capacidad de 256,82 medidas por segundo contra 126,71 de JSON y 26,60 de XML. Si bien Brief no soporta datos complementarios, los números de la simulación permitirían concluir que los datos complementarios tienen un elevado coste para el procesamiento y deben utilizarse con racionalidad. Es decir, por defecto se debiera utilizar Brief en los mensajes y XML/JSON solo cuando se requiera intercambiar datos complementarios. Se introdujeron diversos puntos de contacto entre escenarios, estados y la distancia compuesta dentro de la arquitectura:

- La posibilidad de conocer el estado actual de una entidad y el escenario de contexto en tiempo real, permite utilizar indicadores y criterios de decisión específicos acordes al instante de análisis de datos. Ello produce que el análisis de criterios de decisión sea pertinente a la situación (y no a la definición original).
- La distancia compuesta provee una lista de proyectos ordenados por similitud considerando la perspectiva estructural y comportamental de los restantes proyectos de medición. Esto ordena el espacio de búsqueda y lo limita, dado que permite definir reglas tales como “Solo considerar proyectos con una distancia compuesta >=0.6”.
- De este modo, la búsqueda de recomendaciones mediante la lista acotada de proyectos similares promueve la reutilización de conocimiento y experiencia previa máxime cuando un proyecto no posee conocimiento específico.

Mejorar la pertinencia de los criterios de decisión para el análisis de medidas permite focalizar en recomendaciones acordes a la situación en un instante dado. Ante la ausencia de estas, la distancia compuesta permite localizar proyectos similares para reutilizar conocimiento y experiencias previas.

Capítulo 6	Tecnologías de Soporte a la Arquitectura de Procesamiento

Capítulo 6. Tecnologías de Soporte a la Arquitectura de Procesamiento

Introducción

El capítulo anterior introdujo la arquitectura de procesamiento basada en metadatos de mediciones y su relación con la definición de proyectos de medición. Adicionalmente, describió nuevas propuestas para el intercambio de medidas (Brief), junto con una organización jerárquica de la recolección de datos (adaptadores y pasarelas) con posibilidad de transmisión indirecta. Se ejemplificó el modo en el que estados de entidad y escenarios se calculaban junto con la matriz de ocurrencia para estimar probabilidades empíricas. Se detalló el impacto del formato Brief en la recolección de datos frente a sus antecesores junto con el rol de la distancia compuesta para la estimación de puntajes (recomendaciones) y para guiar la búsqueda de recomendaciones limitando el espacio de búsqueda.

El presente capítulo introduce los detectores de cambio en el adaptador de medición como esquema para optimizar la transmisión de datos y reusar el conocimiento descrito en la definición de proyecto. Se plantean organizaciones de búfer particulares para esta estrategia junto con estimaciones del comportamiento como referencia.

Se introduce el rol del descarte selectivo basado en puntuación Z empleando metadatos de medición en el origen de los datos (es decir, el adaptador). Por un lado, permite descartar datos siguiendo las políticas de prioridad del proyecto cuando la tasa de arribo supera a la de procesamiento. Por otro lado, plantea organizaciones internas del búfer tendientes a mejorar la confiabilidad del recolector. Dado que su uso es opcional en los adaptadores de medición, se proveen patrones de referencia para poder contrastar con los ambientes donde desee utilizarse.

Se describe el uso de árboles de Merkle como registro de integridad. Ello permite almacenar la huella de los datos transmitidos o recibidos (descartando los datos originales) y poder verificar contra el origen su estado (es decir, si han sido de algún modo modificados o no). Se proveen patrones de referencia respecto de las operaciones involucradas.

Finalmente, se describe la implementación distribuida basada en Blockchain del registro unificado de nodos, donde la actualización de este es originada por cada nodo y requiere consenso de los vecinos para su aprobación. Esto permite que el registro unificado no tenga un control central sino dado por la trayectoria de cada nodo y su peso relativo en la recolección de datos. Adicionalmente, se provee una implementación de referencia.

El capítulo se soporta en las siguientes publicaciones efectuadas a lo largo del proceso de investigación:

- Diván, M. Sánchez-Reynoso, (2022). **“Transformations through Blockchain Technology”.** Towards a Distributed Record of Measurement Adapters Powered

by Blockchain Technology. ISBN 978-3-030-93343-2. pp113-135. Dinamarca. Cham. Springer Cham. https://doi.org/10.1007/978-3-030-93344-9_5

- Diván, M. Sánchez Reynoso, M. (2021) "**A Metadata and Z Score-based Load-Shedding Technique in IoT-based Data Collection Systems"**. International Journal of Mathematical, Engineering and Management Sciences. ISSN: 2455-7749. e-ISSN: 2455-7749. Elsevier. Vol.6, nro 1. pp 363 – 382. https://ijmems.in/volumes/volume6/number1/23-IJMEMS-SBS19-34-6-1-363-382-2021.pdf

- Diván, M. Sánchez Reynoso, M. (2021) **"Metadata-based measurements transmission verified by a Mervle Tree"**. Knowledge-Based Systems. ISSN: 0950-7051. Ed. Elsevier Science BV. Vol. 219. pp 1 -17. http://dx.doi.org/10.1016/j.knosys.2021.106871

- Diván, M. Sánchez-Reynoso, M. Gonnet, S. (2021) **"Recent Applications of Federated Learning in Edge and IoT Environments: A Review**". Proceeding of the 2021 5th International Conference on Information Systems and Computer Networks (ISCON). IEEE. https://doi.org/10.1109/ISCON52037.2021.9702466

- Diván, M & Sánchez Reynoso, M (2020) "**Optimizing Data Transmission from IoT devices through Weighted Online Data Changing Detectors**" Advances in Data Science and Adaptive Analysis. ISSN 2424-922X. Vol 12 (2). 2041001 (pp.1:33). https://doi.org/10.1142/S2424922X20410016

- Diván, M & Sánchez Reynoso, M (2020) "**Relocating the Load-Shedding Strategy in the Data Stream Processing Architecture**" In IEEE 2020 Argencon. Resistencia, Chaco. 2 al 4 de diciembre. En Prensa. Se adjunta PDF firmado por IEEE.

- Diván, M & Sánchez Reynoso (2019) "**A Load-Shedding Technique based on the Measurement Project Definition**". In V. Jain, S. Patnaik, F. Popentiu Vladicescu, and I.K. Sethi (Eds.). Proceedings of 5th International Conference on Intelligent Computing, Communication & Devices (ICCD 2018), Xi'an, China, November 22-24 of 2018. In Advances in Intelligent Systems and Computing, Springer Nature Singapore. pp.1027-1033. ISSN 2194-5357. https://doi.org/10.1007/978-981-13-9406-5_122

- Diván, M & Sánchez Reynoso, M (2018) "**The Real-Time Measurement and Evaluation as System Reliability Driver**". Book Chapter in "System Reliability Management: Solutions and Technologies". Anand, A & Ram, M (Eds.). CRC Press,

Taylor & Francis Group. Pp. 161-188. https://doi.org/10.1201/9781351117661-11

6.1 Transmisión de Datos y Detectores de Cambio de Datos

Esta sección describe un acercamiento para detectar cambios en forma ponderada por la definición de proyecto de medición. Se describe el uso de detectores de cambio y barreras temporales junto con una organización de búfer particular que permita informar las medidas más recientes desde el origen (es decir, el adaptador de medición). Esto permite informar medidas pertinentes al monitoreo (es decir, para proveer una prueba de vida o cuando un cambio ha sucedido en la entidad/contexto bajo monitoreo), focalizando en la búsqueda de recomendaciones ante cambios cuantificables. Es decir, la distancia compuesta computa el comportamiento y evolución de las métricas de un proyecto basado en tales cambios para definir la lista de proyectos más similares.

6.1.1 Filtros de datos en línea y Ponderación de las métricas

Los datos que provienen desde los sensores se encuentran bajo la influencia del ruido por diversas razones, tales como las condiciones ambientales, mal funcionamiento, incertidumbre, problemas de calibración, etc. Si bien el ruido podría ser reducido, no puede ser completamente eliminado debido a que las variables del proceso de medición son susceptibles al ruido. Es decir, un sensor funcionando perfectamente funcionando e informando la misma temperatura implicaría una varianza cero, pero es irreal debido a la entropía que engloba al sistema. Por esa razón, la varianza podría tender a cero en dicho caso, aunque no alcanzarla [148]. En este contexto, la idea subyacente del adaptador de medición es atrapar el cambio de datos tan pronto como suceda para enviar el conjunto de datos pendientes que llevaron a que éste suceda, en lugar de estar transmitiendo cada vez que un sensor provee datos. Claro que el punto neurálgico aquí es detectar el instante en que el cambio en el comportamiento de los datos se produce.

En condiciones estables, la medición desde el sensor podría ser analizado como un proceso estacionario, donde la media y la varianza debieran ser invariantes. En caso de que un cambio ocurra en alguno de los mencionados valores, implicaría que las condiciones cambiaron [149]. La idea básica para un detector de cambio de datos es detectar el instante en el cual la media ha cambiado. Sin embargo, debe resaltarse que este cambio debe analizarse en un contexto de flujos de medidas donde el procesamiento de datos es incremental y en línea (es decir, al mismo tiempo en que los datos vienen desde los sensores).

Basado en el método SPC definido en [150], la idea es reaccionar cuando la suma acumulativa (CUSUM) de desviaciones de la media excede un cierto límite. La Ecuación 17 expresa esta idea, indicando con ´N´ el número de datos empleado para el cálculo.

Ecuación 17 Sumas Acumulativas

$$if\ |CUSUM| > 3 * \sqrt{N} = \left| \sum \frac{(x - \bar{x}_{old})}{\hat{\sigma}_x} \right| > 3 * \sqrt{N}$$
$$= \left| \sum (x - \bar{x}_{old}) \right| > 3 * \sqrt{N} * \hat{\sigma}_x$$

Los supuestos alrededor del detector de cambios son los siguientes:

1. El proceso es estacionario (es decir, media y varianza no cambian en el tiempo),
2. El ruido es independiente en cada muestra (es decir, no está correlacionado),
3. CUSUM es una variable aleatoria inicializada en cero (es decir, ninguna variación es detectada al inicio).

La ecuación permite observar que |CUSUM| crece sistemáticamente con cada valor, representando una magnitud acumulativa de desviaciones sobre la media (desde el último cambio detectado).

Siguiendo la idea asociada con las desviaciones acumulativas en un proceso estacionario, en [151], los autores proponen un método incremental y en línea que permite detectar la ocurrencia de cambios de datos, ajustando la nueva media de acuerdo con el proceso que tiende a ser estable. Luego de lo cual, un filtro autoajustable basado en desviaciones acumulativas fue introducido por Cao y Rhinehart [152], incorporando como idea que la nueva media será calculada solo cuando exista suficiente confianza estadística de que la vieja media está desactualizada.

Así, a partir de la Ecuación 17, la primera medida recibida no será $\bar{x}_{old}$ sino que puede calcularse a partir de la segunda medida considerando la primera como $\bar{x}_{old}$. Desde allí, el punto pendiente en la ecuación es cómo calcular incrementalmente la desviación en un contexto de flujos de datos. En [153], Alford plantea la propuesta descrita por la Ecuación 18.

Ecuación 18 Estimación de la Desviación en forma Incremental

$$\hat{\sigma}^2_{f_{new}} = \left(\frac{M-2}{M-1}\right) * \hat{\sigma}^2_{f_{old}} + \left(\frac{1}{M-1}\right) * \frac{1}{2} * (x_{new} - x_{old})^2$$

El primer término de la suma será cero para los dos primeras medidas recibidas dado que no existe varianza detectada aún. Sin embargo, al arribar la segunda medida, el segundo término de la suma comienza a estimar la varianza. ´M´ representa el número de datos utilizados en la estimación de la varianza. En [153], un M=11 es recomendado para mantener un equilibrio entre variabilidad y detección de cambio de datos.

Ecuación 19 Estimación de la Desviación en forma Incremental con M=11

$$\hat{\sigma}^2_{f_{new}} = \frac{9}{10} * \hat{\sigma}^2_{f_{old}} + \frac{1}{20} * (x_{new} - x_{old})^2$$

La
Ecuación 19 surge de reemplazar en la Ecuación 18 M con 11, y su implementación algorítmica es incremental dado que es calculada ante el arribo de cada nueva medida. Adicionalmente, la Ecuación 17 podría utilizar la estimación de la
Ecuación 19 para implementar la detección del cambio de datos. Así, cuando un cambio ocurre, la nueva media puede ser estimada como se indica en la Ecuación 20.

Ecuación 20 Estimación de la Media

$$\bar{x}_{new} = \bar{x}_{old} + \hat{\sigma}_x * \frac{CUSUM}{N}$$

De este modo, mediante la Ecuación 17 y la Ecuación 19 es posible detectar el cambio en línea, mientras que la Ecuación 20 permite estimar la media en tiempo real. No obstante, debe mencionarse que este acercamiento es univariado, mientras que en un proyecto de medición, la detección de cambios e incidencia global no tiene la misma importancia en todas las variables (o métricas) involucradas.

Una implementación de referencia en JAVA es provista mediante el repositorio en GitHub dentro de la librería pabmmCommons. A su vez, la librería incorpora un detector de cambios multihilo para el proyecto de medición, donde la ponderación de cada variable involucrada se considera antes de disparar una señal de transmisión al búfer de datos.

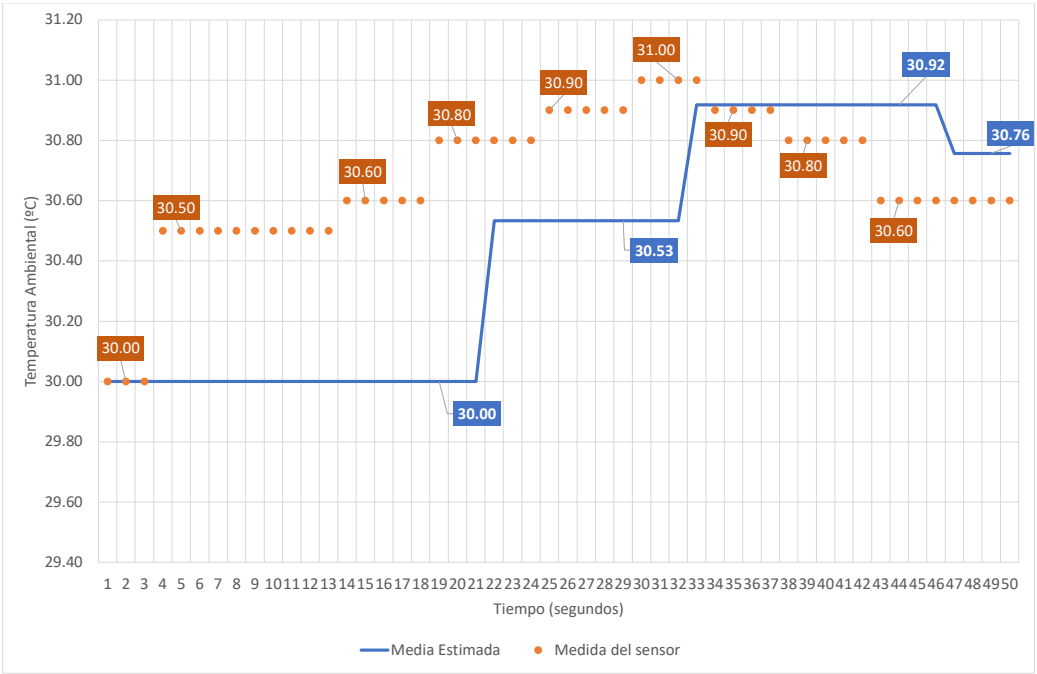

Figura 37 Estimación de la Media utilizando medidas durante 50 segundos

Para ejemplificar la idea, suponga que se leen medidas desde un sensor de temperatura durante 50 segundos (Puntos en la Figura 37). Cada vez que una medida arriba en ese lapso de tiempo, se acumula la diferencia entre el nuevo valor y la media estimada. La línea continua en la Figura 37 representa la media estimada y sus cambios. Notar que esta no cambia con cada nueva medida que arriba, sino que cambia cuando existe cierta cantidad de evidencia desde la serie de datos para la métrica dada. De este modo, la media es actualizada cuando el valor acumulado para CUSUM es superior a $3 * \sqrt{N} * \hat{\sigma}_x$ utilizando la

Ecuación 19 para estimar la varianza. De este modo, la media estimada se actualiza utilizando la Ecuación 20, mientras que cada actualización representa un incremento en el contador de la métrica para saber que la serie de datos cambió un cierto número de veces desde la última transmisión de datos.

Como puede apreciarse en la figura anterior, la media estimada se actualizó cuatro veces en 50 segundos (es decir, 30ºC, 30,53ºC, 30,92ºC, y 30,76ºC), ello implica que el contador acumulado para la temperatura ambiental será 4 luego de 50 segundos (los 4 cambios).

Es importante mencionar que los cuatro cambios no implican que existen cuatro transmisiones de datos, dado que el cambio es para una métrica que puede tener ponderaciones diferentes de otras en el proyecto de medición. Aquí es donde se articulan los detectores de cambios con la ponderación de las métricas del proyecto de medición.

Suponga que se desea monitorear principalmente la humedad del suelo y por ello se le asigna una ponderación de 0,6. Por otro lado, la temperatura y humedad ambiental se les asigna una ponderación de 0,2 respectivamente. Los contadores que representan cambios en las estimaciones de las medias son definido como $q_{smoisture}$, $q_{temperature}$, y $q_{humidity}$. Para que se produzca la transmisión de datos desde el adaptador de medición, la suma ponderada de las métricas debe exceder un umbral arbitrario definido como parámetro del filtro.

Tabla 28 Efecto de la Suma Ponderada de las Métricas en la Transmisión de Datos

Tiempo (Segundos)	$q_{moisture}$	$q_{temperature}$	$q_{humidity}$	$0,6*q_{moisture}+ 0,2*q_{temperature}+ 0,2*q_{humidity}$
1	0	1	0	0.6*0 + 0.2*1 + 0.2*0= 0.2
22	0	2	0	0.6*0 + 0.2*2 + 0.2*0= 0.4
33	0	3	0	0.6*0 + 0.2*3 + 0.2*0= 0.6
47	0	4	0	0.6*0 + 0.2*4 + 0.2*0= 0.8
55	1	4	0	0.6*1 + 0.2*4 + 0.2*0= 1.4
62 (*)	1	4	1	**0.6*1 + 0.2*4 + 0.2*1= 1.6**
63	0	0	0	0.6*0 + 0.2*0 + 0.2*0= 0.0

Como se puede observar la Tabla 28, la sola ocurrencia del cambio de la media estimada para una métrica dada no es suficiente para producir la transmisión de datos.

Suponga que los únicos cambios durante los 50 segundos ocurren en la temperatura ambiental, mientras que la humedad del suelo y humedad ambiental informarán cambios a los 55 y 62 segundos respectivamente. De este modo, dado un umbral arbitrario de 1.5, la transmisión de datos siguiendo la anterior tabla ocurriría en el segundo 62 debido a que en dicho instante la expresión ponderada conjunta excede el umbral arbitrario establecido (1.5). Luego de ello, todos los contadores se reinician esperando por nuevos cambios de datos que permitan incrementar los contadores.

6.1.2 Organización Dinámica del Buffer para Soportar Detectores de Cambio de Datos

Para soportar los detectores de cambios en línea, la Figura 38 describe una organización alineada con la definición del proyecto de medición. Se organiza la correspondencia entre métricas y atributos junto con la ventana lógica y los detectores de cambio.

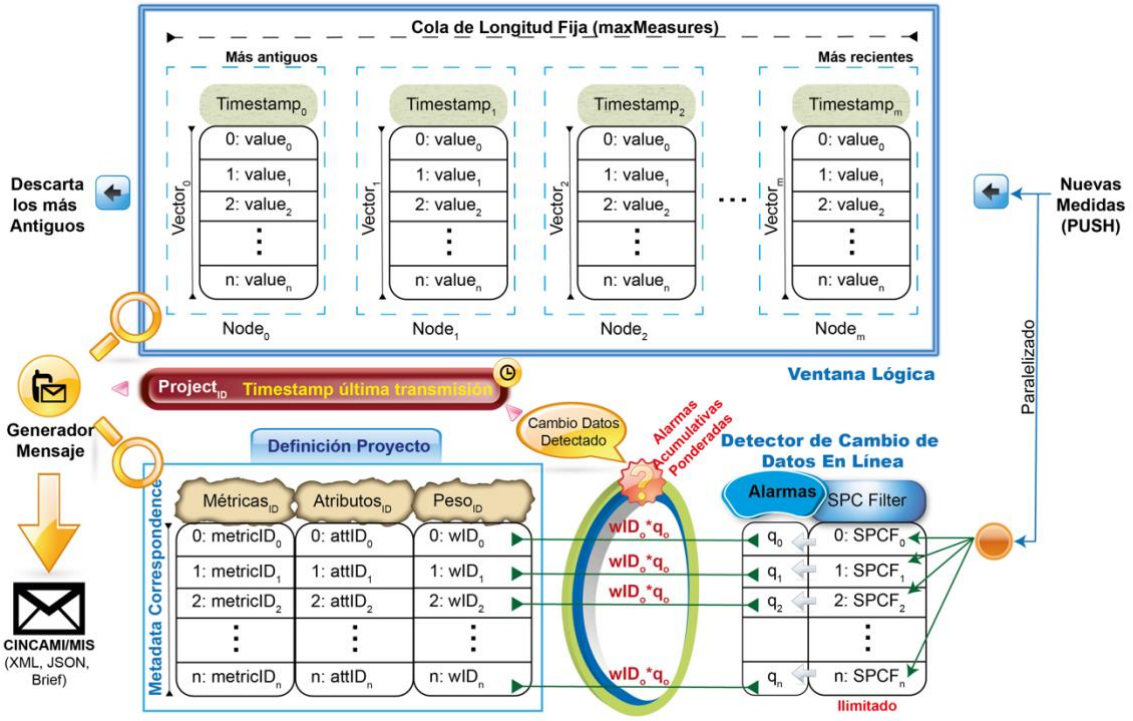

Figura 38 Articulación del Búfer de Datos y los Detectores de Cambio

Inicialmente, el identificador de proyectos ($project_{ID}$) se mantiene en memoria junto con las métricas y atributos asociados. Desde allí, se crea una matriz bidimensional que contiene tres columnas: el ID de métrica, ID de atributo y la ponderación asociada. La matriz contendrá tantas filas como métricas se definan en el proyecto.

Como se puede apreciar, el detector de cambio de datos del proyecto se compone de un conjunto de detectores de cambios individuales a nivel de métrica junto con su acumulador respectivo. El acumulador indica las alarmas disparadas por su detector desde la última transmisión de datos.

La ventana lógica se organiza como una cola ordenada de nodos que contienen el conjunto de valores para las métricas en un instante dado (timestamp) y respetando un orden estricto dado por las métricas. Por ejemplo, si la primera métrica es la temperatura ambiental, las medidas en la primera posición de la cola para un instante de tiempo será siempre la temperatura ambiental. Cuando nuevas medidas de los sensores arriban, ellas son agregadas al final de la cola desplazando las medidas más antiguas. Es decir, si la incorporación de las nuevas medidas excediere el tamaño máximo de la ventana lógica, se descartarían las medidas más antiguas (extremo izquierdo de la ventana lógica en la Figura 38) para permitir el ingreso de las más recientes (extremo derecho de la ventana lógica en la Figura 38). De este modo, se evita un desborde de buffer cuando no hay transmisión de datos y se intenta exceder la capacidad de la ventana lógica.

Los detectores de cambio en línea estiman la media y varianza para cada métrica en forma continua e incrementalmente junto con el arribo de las medidas, motivo por el cual no necesitan almacenar más datos que la estimación. Cuando un detector para una métrica detecta el instante de cambio, incrementa su contador asociado informándolo al detector de cambio del proyecto. Este ejecuta la suma ponderada contemplando el nuevo estado de los acumuladores para determinar si se excede o no el umbral establecido en el proyecto. De excederse, el detector de cambios informa al adaptador de medición que debe generarse el mensaje (puede ser como JSON, XML, o Brief) y su transmisión correspondiente. Luego de la transmisión de datos, se limpia el búfer de datos, la marca de tiempo de la última transmisión se actualiza y los contadores de las métricas se reinician.

Una implementación de referencia pen JAVA para el búfer de datos y los detectores de cambio se provee bajo los términos de la licencia Apache 2.0 en GitHub dentro de la librería pabmmCommons.

6.1.3 Estimando el Comportamiento de los Detectores de Cambio de Datos

Se realizaron diferentes simulaciones para obtener una referencia sobre el comportamiento de los detectores de cambio, filtros individuales, y búfer de datos. Las mismas se ejecutaron sobre una MacBook Pro con MacOS Catalina 10.15.4 con 16GB RAM sobre Java 8. El código fuente para reproducir las simulaciones están disponibles dentro de la librería pabmmCommons. Las simulaciones ejecutadas fueron:

1. **Simulación 1:** Analizar los tiempos de creación de los detectores junto con sus tamaños asociados. Se empleó instrumentación mediante agentes de JAVA para medir los tamaños de memoria. Se varió el número de métricas entre 1 y 100 y por cada iteración, se crearon los detectores para las métricas indicadas. Se midió el tiempo de creación junto con el tamaño del detector previamente y posteriormente a que las medidas se incorporaron.

2. **Simulación 2**: Se analizó el comportamiento del detector de cambio de datos, considerando el número de alarmas disparadas durante cinco minutos continuos. Se varió el número de métricas de 10 a 30, mientras que el umbral (parámetro TRIGGER) varió entre 2 y 3. Para cada combinación del número de métrica y umbral, se creó el respectivo detector de cambio de datos dinámicamente y se analizó el número de alarmas disparadas por cinco minutos. Las medidas se generaron arbitrariamente y en forma aleatoria con una media de 7 y una desviación de 2.

3. **Simulación 3:** Similar a la anterior pero en donde las medidas informadas tenían una media de 7 y una desviación de 3.

4. **Simulación 4:** Se focalizó en medir las operaciones individuales ejecutadas por el filtro mientras se le proveían medidas aleatorias. Es decir, se focalizó en estimar el coste de *addMeasure* (al agregar una nueva medida) y *compute* (estimar la media y varianza incrementalmente).

5. **Simulación 5:** Se centró en analizar el tamaño del búfer de datos durante 5 minutos mientras arriban medidas. Se tomó como referencia un búfer para 10 métricas y se le proveyó datos cada 100 ms por 5 minutos. El número máximo de medidas a mantener en memoria se definió en 1000 y el parámetro *maxTolerance* se fijó en 5 (umbral de la suma ponderada de cambios del proyecto). Se midió el consumo de memoria asociado.

6. **Simulación 6:** Se midió el comportamiento del bufer junto con los detectores de cambio, empleando barreras temporales. Se proveyeron medidas con media 9 y desviación 2. Se creó un buffer para 10 métricas con igual ponderación, activando la barrera temporal cada 15 segundos. El parámetro *maxTolerance* se fijó en 5 y *maxMeasures* en 1000. El búfer recibió medidas cada 100 ms por 5 minutos, tiempo en el cual se midió las alarmas disparadas y tipos (temporal o cambio), y el número de medidas informadas en la transmisión.

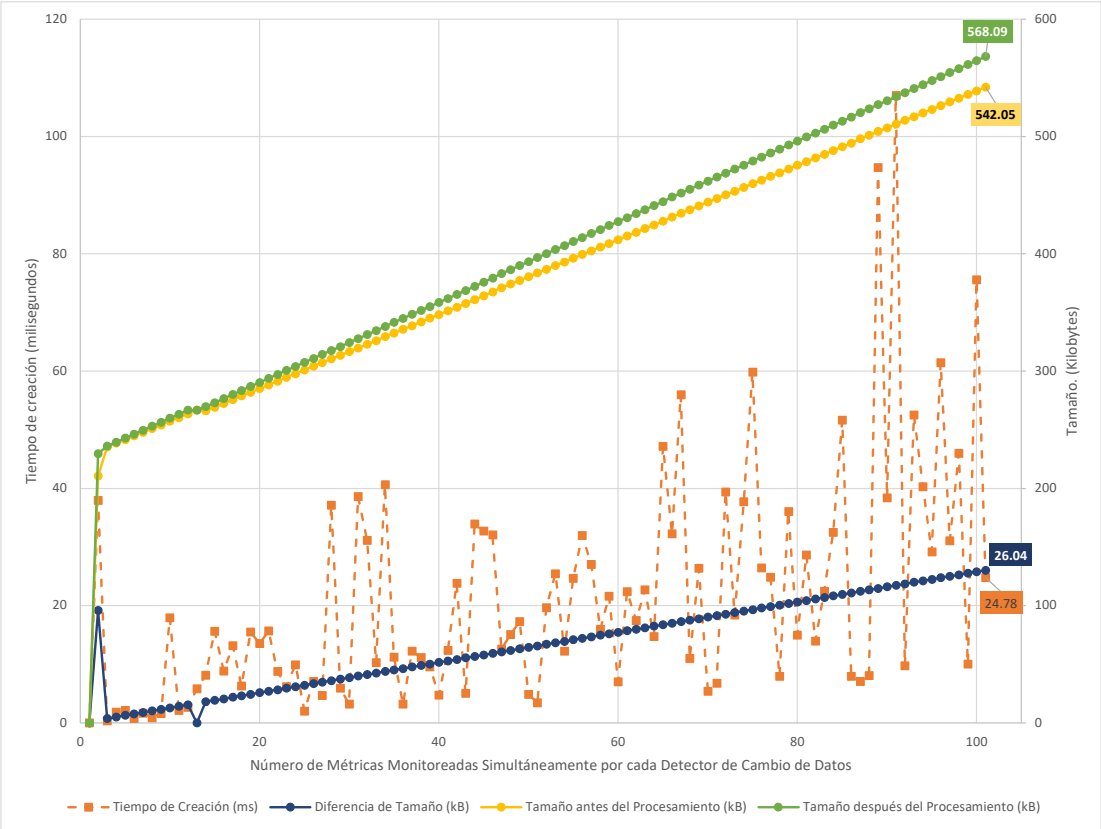

Figura 39 Tamaño consumido y tiempo en la creación y procesamiento de detectores de cambio en línea

La Figura 39 sintetiza los resultados de la primera simulación para los tiempos de creación de un detector en línea mientras la cantidad de métricas a monitorear varía juntamente con el tamaño consumido antes y después del procesamiento de una medida dada.

La serie de tiempo para la creación (línea punteada, utilizando el eje ordenado izquierdo) representa el tiempo en milisegundos requerido para crear el detector de cambio de datos dependiendo del número de métricas monitoreadas. Como se puede apreciar, el peor tiempo es inferior a 120 ms, mientras que un tiempo optimista indicaría que en 24.78 ms sería posible crear un detector para el monitoreo de 100 métricas. La línea continua, utilizando el eje ordenado derecho india el tamaño relacionado al detector. Es decir, para monitorear 100 métricas simultáneamente se consumiría 542,05 kB antes de procesar datos y 568,09 kB con las estimaciones incorporadas (luego de procesar datos). Así, los datos relacionados con las estimaciones consumirían 26,04 kB en 100 métricas.

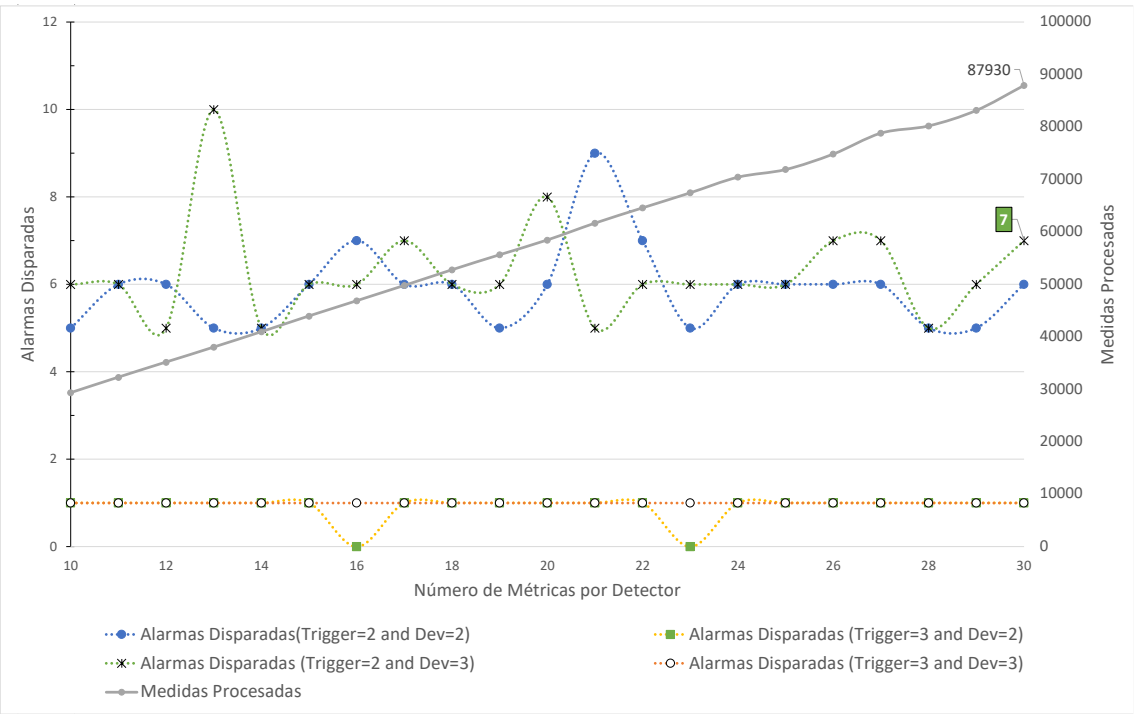

Figura 40 Tamaño y Tiempo Consumido en la Creación y Procesamiento de un Detector de Cambio de Datos

La Figura 40 describe la evolución de los datos procesados mientras el número de métricas se incrementa (línea continua con referencia al eje ordenado derecho). Tales resultados surgen del procesamiento continuo de datos en las simulaciones 2 y 3. Vale la pena destacar la diferencia en el detector de cambio de datos cuando TRIGGER y los datos recibidos varían (Líneas punteadas con referencia al eje ordenado izquierdo). Cuando TRIGGER se fijó en 3, el número de alarmas disparadas no excedió 1. Sin embargo, cuando TRIGGER se fijó en 2, el detector de cambios es más sensible a disparar alarmas y limita el rango de variación tolerado. El peor escenario visualizado correspondió a 10 alarmas (y por ende transmisiones) a lo largo de 5 minutos, es decir, 1 transmisión cada 30 segundos . Si ello se contrasta contra 1 transmisión por segundo, se estaría decidiendo entre transmitir 10 veces o 300 veces (5 minutos *60 segundos) sin pérdida de datos significativa. La tasa de procesamiento unitario en las simulaciones 2 y 3 varió establemente entre 9,56 y 9,71 medidas en cada métrica por segundo (alrededor de 291,3 medidas por segundo en 30 métricas).

La Tabla 29 indica una serie de medidas descriptivas relacionadas con la simulación, focalizada en los tiempos de operación para agregar una medida y respecto de la estimación de la media y varianza. Como es posible apreciar, la operación *addMeasure* es la más costosa que la estimación de media y varianza, posiblemente debido a la necesidad de implementar operaciones thread-safe para incorporar las medidas.

Tabla 29 Perspectiva Comparativa de los Tiempos de Operación para addMeasure y Compute

Concepto	addMeasure (ms)	Compute (ms)
Mínimo	0,000178	0,000081
1er Cuartil	0,000684	0,000163
Mediana	0,000918	0,000238
Media	0,002343	0,001246
Media Recortada (5%)	0,001145	0,000528
3er Cuartil	0,001405	0,000478
Máximo	1,880212	0,268523

Ambas operaciones sufren de valores extremos, aspecto que puede visualizarse contrastando la media, mediana y media recortada. De este modo, la mediana para la operación de determinar la presencia (o no) de cambios consumió 238 ns (0,000238 ms), mientras que la mediana para la operación de agregación de medidas requirió 918 ns (0,000918 ms).

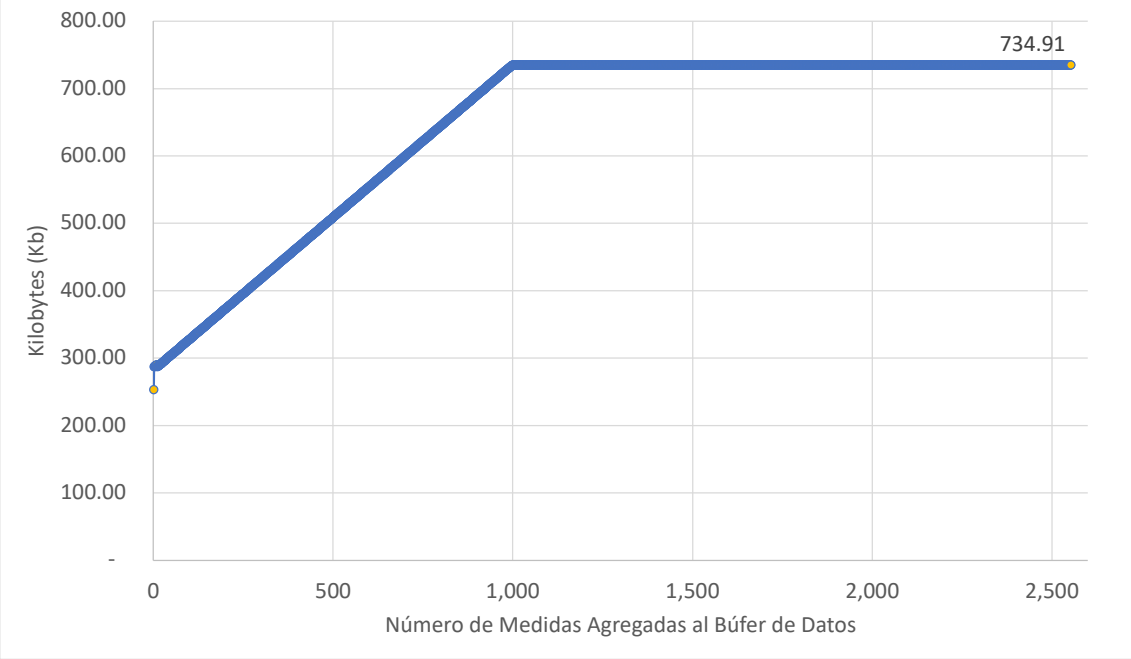

Figura 41 Evolución del tamaño del Búfer para una capacidad máxima de 1000 medidas

La Figura 41 sintetiza los resultados de la simulación 5 cuando el búfer constituido como ventana lógica (con capacidad máxima de 1000 medidas) es poblado con medidas durante 5 minutos. Se le proveyeron poco más de 2500 medidas y como puede apreciarse, creció hasta alcanzar su límite máximo, momento en que se manejó en forma circular descartando las viejas medidas para dar lugar a las nuevas lo que mantuvo estable el tamaño final en 734,91 KB.

La Figura 42 resume el comportamiento conjunto manifestado por el búfer de datos en conjunto con la barrera temporal y los detectores de cambio. Las barreras temporales implican que al menos habrá una transmisión de medidas cada cierta periodicidad para

dar prueba de vida. Sin embargo, de existir una transmisión derivada por el cambio en los datos, el contador de la barrera temporal se reinicia.

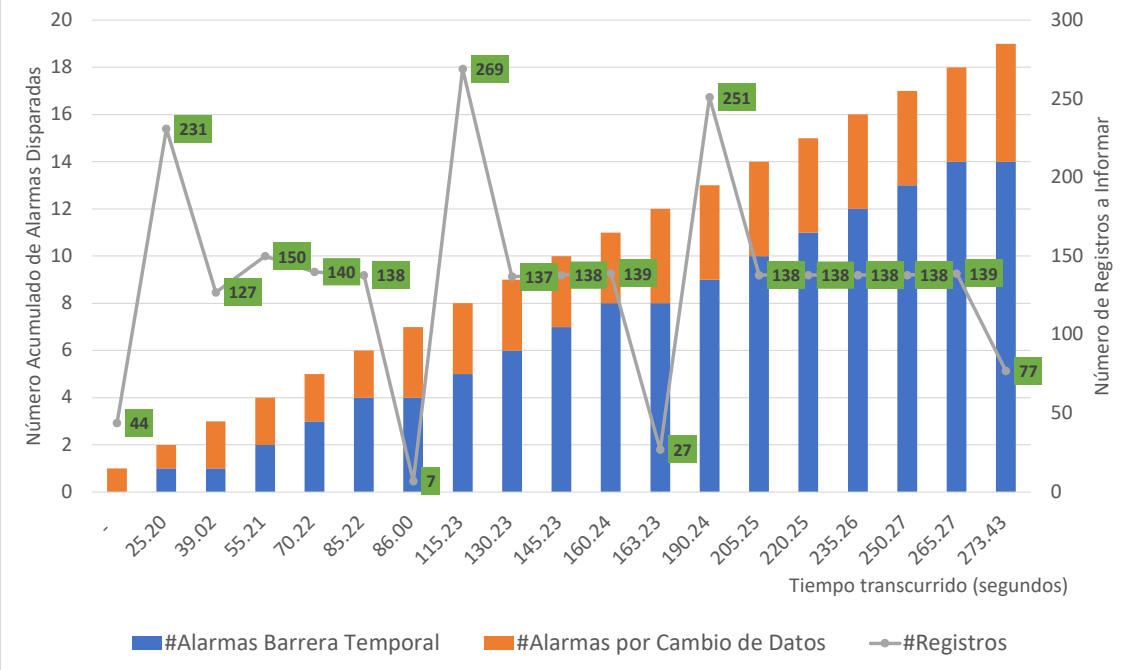

Figura 42 Comportamiento conjunto del búfer, la barrera temporal, y los detectores de cambio

Cada categoría en el eje de las abscisas representa el instante en el que una alarma es disparada. Como se puede apreciar, la barrera temporal es la razón más frecuente contrastado con aquellas originadas por los cambios en los datos, alcanzando 14 transmisiones de un total de 19 (73,68%). Solo 5 de 19 correspondieron a alarmas que se originan por cambios de datos (26,32%). La estrategia articulada entre barreras temporales y detectores de cambios es una mejor opción que transmitir medidas cada segundo. Es decir, serían 19 transmisiones versus 300 requeridas en 5 minutos (Solo el 6,33%). Los números resaltados de la figura representan la cantidad de registros a transmitir (sea por la barrera temporal o porque generaron el cambio en los datos), para la última alarma se indicaría que el mensaje contendrá 77 registros, que a 10 métricas por registro implicarían 770 medidas.

Este control en la fuente de datos le permite a la distancia compuesta contar con datos que se asocian con cambios en los comportamientos de sus métricas (con la suficiente prueba estadística), evitando procesar aquellos reiterativos que no aportan demasiado para explicar la serie de datos de un atributo o propiedad contextual. Ergo, se genera un efecto positivo en el cómputo de similitudes comportamentales entre proyectos, a la vez que se minimiza el intercambio de datos a lo necesario.

6.2 Descarte Selectivo basado en Z-Score y Metadatos de Medición

Esta sección describe la articulación del búfer de datos para la incorporación de descarte selectivo basado en puntuaciones Z y ponderadas por la importancia relativa

de sus métricas. La primera parte aborda la estimación incremental y el impacto de la ventana lógica en el búfer. Seguido, se describe la técnica de descarte. Finalmente, se proveen una serie de simulaciones discretas como patrón de referencia de aplicabilidad.

6.2.1 Búfer de Datos y Estimación Incremental

A partir de la definición de proyectos intercambiada en los nodos y capas de procesamiento en la nube (Ver BriefPD en Capítulo 3), es posible lograr diferentes organizaciones del búfer de datos en los adaptadores de mediciones que sirvan para ajustar diferentes escenarios, por ejemplo, de recursos y capacidad de almacenamiento limitada en un nodo.

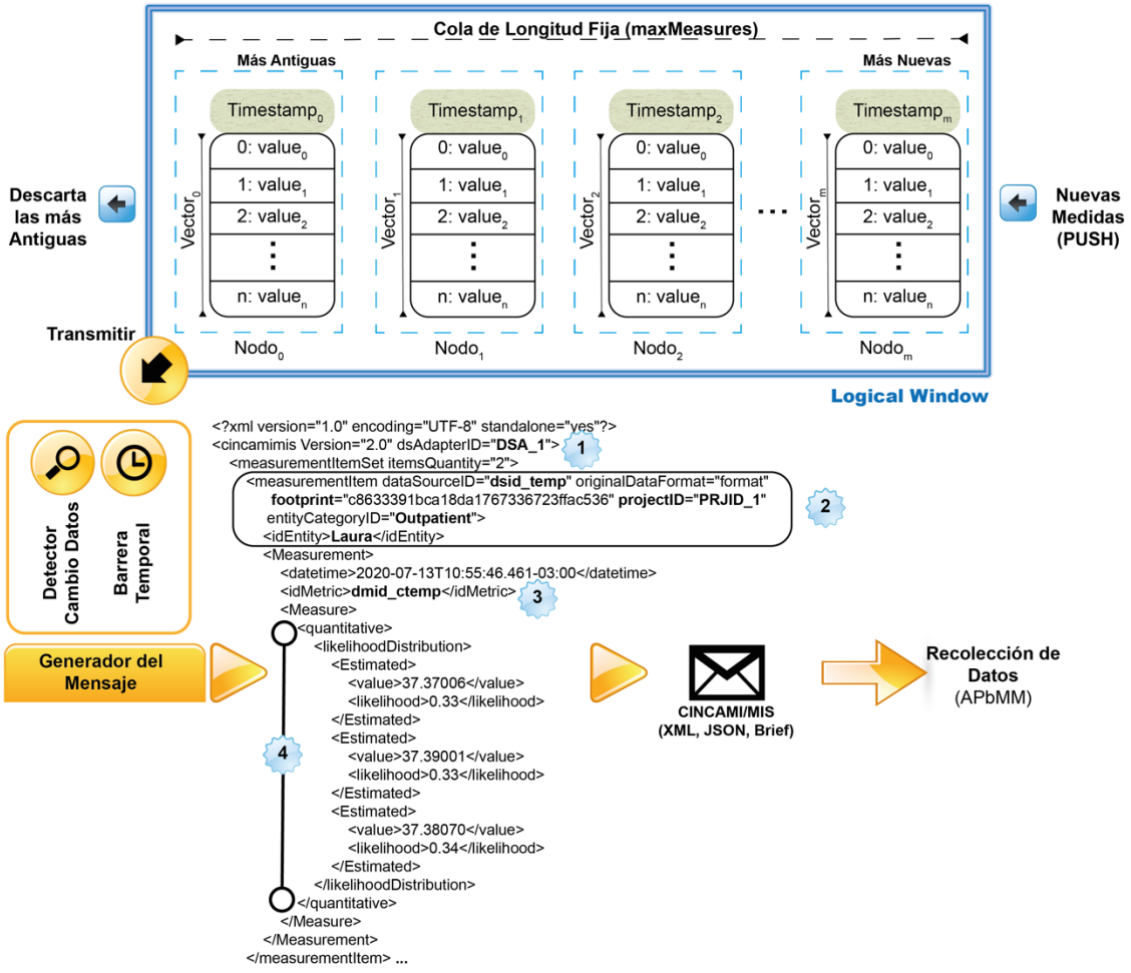

Figura 43 Perspectiva Conceptual de la Organización del Búfer de Datos alineado con el Esquema de Intercambio de Mediciones

La Figura 43 sintetiza la organización y comportamiento global del búfer de datos para soportar descarte selectivo y detectores de cambio en el adaptador de medición introducidos en la Figura 38. Estas organizaciones se crean dinámicamente en base a la configuración del nodo, por ejemplo, si desea o no emplear detectores de cambio, barreras temporales, entre otras funcionalidades [134]. El descarte selectivo consiste en

retener las medidas asociadas con aquellas métricas de mayor importancia para el proyecto de medición cuando la capacidad de procesamiento se ve comprometida. Dado que esta importancia relativa para las métricas es fijada por el director de proyecto en base a la importancia de la métrica para el monitoreo del contexto o entidad, posee un impacto directo en la distancia compuesta al momento de calcular la similitud o no en términos de comportamiento esperado. En otras palabras, si se debe calcular la distancia compuesta es lógico que se priorice las medidas más importante. Por ejemplo, siguiendo el ejemplo de la humedad del suelo de la Tabla 28, las medidas a informar prioritariamente serían 1º) Humedad del Suelo (0,6) seguido por la temperatura ambiental o la humedad ambiental indistintamente dado que ambas tienen una ponderación de 0,2.

Como se introdujo en la sección previa, la transmisión ocurre cuando el adaptador recibe la alarma desde la barrera temporal o del detector de cambio de datos. A partir de dicho momento, se genera el mensaje con los datos del búfer. La sección 5.2.1 describió la organización del esquema de intercambio de mediciones y cómo los metadatos se embeben junto con las medidas basados en la definición del proyecto de medición. Por ejemplo, Un mensaje CINCAMIMIS en XML parcial es descripto en la Figura 43. Se indica el adaptador de medición que actúa como traductor entre las medidas planas de los sensores y la generación del mensaje (Estrella con un 1, etiqueta *dsAdapterID*), la fuente de las medidas (etiqueta *dataSourceID* en rectángulo con estrella 2), el proyecto de medición al que pertenecen (etiqueta *projectID* en rectángulo con estrella 2), la categoría de entidad (etiqueta *entityCategoryID* en rectángulo con estrella 2), la entidad en particular con el que se asocian las medidas (etiqueta *idEntity* en rectángulo con estrella 2), la métrica asociada con las medidas (etiqueta *idMetric* en estrella 3), la/s medida/s deterministas o no (estrella 4 se asocia con medidas estimadas). Así, cuando la recolección de datos de PAbMM recibe los flujos de medidas puede procesarlos guiado por la semántica de sus metadatos.

Anteriormente se introdujo la estimación incremental de la media y desviación para los detectores de cambio, aunque tales estimaciones no se limitaban solo a los datos del búfer sino a todos los que habían procesados. Ahora, si se deseare tener una aproximación de la media y desviación incremental limitado a los datos del búfer, debiera considerarse no solo la incorporación de nuevas medidas sino también el momento en que se descartan las antiguas. Esto es importante porque si se desea obtener una puntuación Z se requieren media y desviación para estandarizar sus valores evitando el efecto de los valores atípicos (outliers) tal y como expone la Ecuación 21.

Ecuación 21 Fórmula para la Puntuación Z

$$Z_{score} = \frac{x - \bar{x}}{s}$$

Dado que se reciben medidas continuamente, debe abordarse una estrategia incremental de estimación de la media aritmética como expone la Ecuación 22 para un tiempo "t" a partir de la estimación del tiempo "t-1".

Ecuación 22 Fórmula de Cálculo Incremental de la Media Aritmética

$$\bar{x}_t = \frac{\bar{x}_{t-1} * n_{t-1} + x_t}{n_{t-1} + 1}$$

La desviación estándar es estimada a partir de la varianza muestral utilizando la Ecuación 23. En este sentido, es importante mencionar que la estimación emplea la media aritmética actual disponible al momento en que la medida arriba.

Ecuación 23 Fórmula de Cálculo Incremental de la Varianza Muestral

$$s_t = \sqrt{\frac{\sum_{t=1}^{n}(x_t - \bar{x}_t)^2}{n-1}}$$

La Ecuación 24 esquematiza dicha situación, donde la media aritmética por cada instante no es necesariamente igual a su predecesora o sucesora debido a la actualización asociada con cada nueva medida arribada.

Ecuación 24 Ejemplo del Cálculo Incremental de la Varianza Muestral

$$s_4 = \sqrt{\frac{(x_1 - \bar{x}_1)^2 + (x_2 - \bar{x}_2)^2 + (x_3 - \bar{x}_3)^2 + (x_4 - \bar{x}_4)^2}{4-1}}$$

La implementación de la Ecuación 23 y la Ecuación 24 requieren mantener en memoria los acumuladores junto con el número de medidas procesadas. Sin embargo, tal y como se introdujo en la sección anterior, el búfer de datos se comporta como una ventana lógica descartando las medidas antiguas ante el arribo de las nuevas cuando su capacidad está completa. Ello afectaría las sumas de las ecuaciones mencionadas dado que parte de tales datos sería descartados. Para evitar esta situación la Ecuación 25 y la Ecuación 26 se ajustaron siguiendo el supuesto que el parámetro *maxMeasures* (cantidad máxima de slots en el búfer de datos) se fija arbitrariamente para este ejemplo en 100.

Ecuación 25 Cálculo Incremental de la Media Muestral con Descarte de Medidas

$$\bar{x}_t = \frac{\bar{x}_{t-1} * n_{t-1} + x_t - x_{t-100}}{n_{t-1} + 1 - 1}$$

La ecuación de la media muestral sustrae el valor descartado (es decir, x_{t-100}), y decrementa en 1 el denominador, al tiempo que el nuevo valor se incorpora (es decir, x_t).

Ecuación 26. Cálculo Incremental de la Varianza Muestral con Descarte de Medidas

$$s_t = \sqrt{\frac{(\sum_{t=1}^{n}(x_t - \bar{x_t})^2) - (x_{t-100} - \bar{x}_{t-100})^2}{n-1-1}}$$

En la Ecuación 26 sucede algo similar a la Ecuación 25. La diferencia relacionada con el vector de medidas más antiguas a ser descartada (es decir, $(x_{t-100} - \bar{x}_{t-100})$) es decrementada del acumulador al mismo instante que en la nueva diferencia se incorpora. El denominador se decrementa en 1 mientras que ´n´es incrementado en paralelo debido a la nueva medida, teniendo siempre como límite superior el parámetro *maxMeasures* (es decir, $n \leq maxMeasures$). La implementación de esta última ecuación requiere que para cada vector de medidas en memoria, se posea la media estimada en cada instante alineado con el proceso de descarte.

Ecuación 27 Alternativa para el Cálculo Incremental de la Varianza Muestral

$$s_t = \sqrt{\frac{n}{n+1} * \left[s_{t-1}^2 + \frac{(x_t - \bar{x}_{t-1})^2}{n+1}\right]}$$

La Ecuación 27 plantea una alternativa al cálculo incremental de la varianza muestral sin necesidad de almacenar medidas en memoria como sugiere la Ecuación 26. Esta alternativa requiere conocer las estimaciones previas de la desviación y media aritmética junto con el número de observaciones. Por un lado, la Ecuación 26 va progresivamente ajustando su valor de acuerdo a los últimos *maxMeasures* datos en consonancia con la estimación de la media, permitiendo una mejor caracterización de este. Por otro lado, la Ecuación 27 estima la varianza desde el principio de la serie de datos y no limitado a los últimos *maxMeasures* valores. De este modo, la Ecuación 27 sería una mejor alternativa para estudiante la estimación de la desviación global para la serie de datos cuando sea requerido. Sin embargo, la Ecuación 26 sería una mejor opción cuando se requiere emplear una suma parcial para estimar la covarianza entre métricas, dado que la desviación estimada debiera ser explicativa de los datos contenidos en el búfer de datos.

Sean “i” y “j” dos métrica a ser implementadas en el adaptador de medición a través de sus sensores, la covarianza entre ellos podría ser calculada basado en la Ecuación 26 como se expone en la Ecuación 28.

Ecuación 28 Cálculo de la Covarianza Muestral Incremental

$$Cov(i,j) = \frac{\left[\sum(x_i - \bar{x}_i) * (x_j - \bar{x}_j)\right] - \left[(x_{i-100} - \bar{x}_{i-100}) * (x_{j-100} - \bar{x}_{j-100})\right]}{n - 1 - 1}$$

La Ecuación 28 requeriría mantener en memoria las diferencias para sustraerlas cuando los vectores de datos son descartados del búfer. De este modo, dado un número de métricas "m" (en el proyecto de medición) y utilizando la anterior ecuación, es posible estimar una matriz de covarianza incremental de acuerdo a la ventana lógica del búfer como indica la Ecuación 29.

Ecuación 29 Cálculo de la Matriz de Covarianza Muestral Incremental

$$Cov = \begin{bmatrix} {s^2}_1 & \dots & Cov_{1m} \\ \dots & \dots & \dots \\ Cov_{m1} & \dots & {s^2}_m \end{bmatrix}$$

La matriz de covarianza descripta en la Ecuación 29 es triangular debido a que $Cov(i,j) = Cov(j,i)$. Ello permite implementar un arreglo unidimensional mapeando los elementos de acuerdo con una matriz triangular, optimizando el uso de memoria. Adicionalmente, utilizando la matriz anterior es posible estimar la correlación de Pearson en forma incremental como indica la Ecuación 30.

Ecuación 30 Cálculo de la Correlación de Pearson Incremental

$$\begin{cases} \forall i \neq j : r_{i,j} = \dfrac{Cov(i,j)}{\sqrt{Cov(i,i)} * \sqrt{Cov(j,j)}} = \dfrac{Cov(i,j)}{s_i * s_j} \\ \forall i = j : r_{i,j} = \dfrac{Cov(i,j)}{\sqrt{Cov(i,i)} * \sqrt{Cov(j,j)}} = \dfrac{{s^2}_i}{s_i * s_i} = 1 \end{cases}$$

La librería pabmmCommons disponible en GitHub bajo los términos de la licencia Apache 2.0, contiene una implementación de referencia para las fórmulas aquí descritas junto con la articulación correspondiente con el búfer de datos.

6.2.2 Técnica de Descarte Selectivo basada en Z-score

Ahora bien, hasta aquí se tiene cómo estimar media, desviación, matriz de covarianza y correlaciones en forma incremental articulándolo con la ventana lógica implementada por el búfer de datos. Esta sección articula dichos conceptos con la puntuación Z y el descarte selectivo de medidas.

Originalmente, el descarte selectivo solo estaba disponible en la recolección de datos en la PAbMM, quien recibe el flujo de medidas desde adaptadores de medidas y

pasarelas. Sin embargo, la idea es complementar a la recolección de datos de PAbMM permitiendo que los adaptadores de medición incorporen descarte selectivo de datos (en inglés, load shedding) antes de que la transmisión suceda. Esto permitiría ahorra procesamiento a PAbMM al tiempo que el adaptador evita transmisiones innecesarias.

La activación del descarte selectivo es un elemento opcional a nivel de adaptador, es decir, el adaptador podría transmitir todas sus medidas si así lo deseare (según su configuración). No obstante, cuando la recolección de datos en PAbMM está cerca de alcanzar su límite de procesamiento, podría solicitar a los adaptadores de medición que activen su funcionalidad de descarte para colaborar en la regulación de la tasa de procesamiento. Un adaptador puede decidir omitir el pedido, aunque sería probable que sus medidas sean descartadas (dependerá de la priorización de las métricas), lo que implicó consumir recursos para transmitir a sabiendas del potencial resultado.

Ecuación 31 Cálculo de la Puntuación para Evaluar la Aceptación de Datos

$$\sum_{i=1}^{m} \vec{z}_{x_i} * \vec{w}_i < acceptanceThreshold$$

La Ecuación 31 permite al adaptador de medición filtrar las medidas a ser retenidas en el búfer. Cuando el descarte selectivo se activa en el adaptador, las transmisiones solo suceden ante la detección de cambio en los datos sin considerar las alarmas por barreras temporales. El vector $\vec{z}_{x_i}$ es el vector de puntuaciones calculado utilizando las estimaciones de la media y desviación muestral para cada métrica. El vector $\vec{w}_i$ contiene las ponderaciones de cada métrica definida en la definición de proyecto. Así, cuando el descarte selectivo está activado, las medidas conjuntas desde los sensores se retienen en el búfer de datos solo cuando el producto no excede el umbral de variabilidad establecido (*acceptanceThreshold*), de otro modo, las medidas son descartadas dado que el foco se aplica al comportamiento general de las series de datos y no a casos atípicos particulares.

Este tipo de estrategia de filtrado es dinámica dado que las estimaciones se actualizan continuamente sobre cada nueva medida que se recibe, y el hecho de descartar variaciones excesivas (por encima de un umbral establecido) se centra en estudiar el comportamiento general y no particular a un caso. Esto impacta directo en la distancia compuesta al momento de localizar proyectos similares desde la perspectiva comportamental.

Adicionalmente, la posibilidad de incorporar matriz de covarianza y cálculo de correlaciones permite localmente estimar asociaciones entre métricas, estén asociadas con la entidad bajo análisis o su contexto (ejemplo, la temperatura ambiental respecto de la frecuencia cardíaca del paciente).

6.2.3 Simulación Discreta del Búfer de Datos

Dado que los adaptadores de medición se asocian con configuraciones de hardware austeras, se llevaron adelante una serie de simulaciones discretas a los efectos de contar con patrones de referencia para la aplicación del descarte selectivo.

Las simulaciones 1 y 2 se focalizaron en medir la operación de agregación de medidas en el buffer cuando el descarte selectivo estaba desactivado (simulación 1) versus activado (simulación 2). Ambas simulaciones se ejecutaron a lo largo de 5 minutos, incorporando medidas aleatoriamente (con media 35 y desviación 2) para 15 métricas cada 100 ms. El búfer de datos se definió con capacidad máxima de 1000 slots. Se midió tanto el tamaño del buffer como los tiempos de la operación de agregado.

Las simulaciones 3, 4, y 5 se orientaron a analizar el comportamiento de las alarmas disparadas para la transmisión de datos.

- La simulación 3 analizó la situación con descarte selectivo desactivado, efectuando transmisiones a través de las alarmas de barrera temporal y cambio de datos.
- La simulación 4 detalla el comportamiento con la técnica de descarte selectivo activado y desactivando las alarmas de barrera temporal. Es decir, solo se transmite cuando sucede un cambio en los datos. Por este motivo, el tiempo de simulación se incrementó a 15 minutos para monitorear 10 métricas concurrentes, manteniendo la ventana lógica del búfer en 1000 slots.
- La simulación 5 expone un escenario donde el descarte selectivo y las alarmas por barrera temporal están desactivadas, mientras que las transmisiones solo suceden por alarmas de cambio en los datos.

La implementación de las simulaciones se encuentran disponibles en la clase mainMeansBufferDetectors class dentro del paquete org.ciedayap.simula package en la librería pabmmCommons.

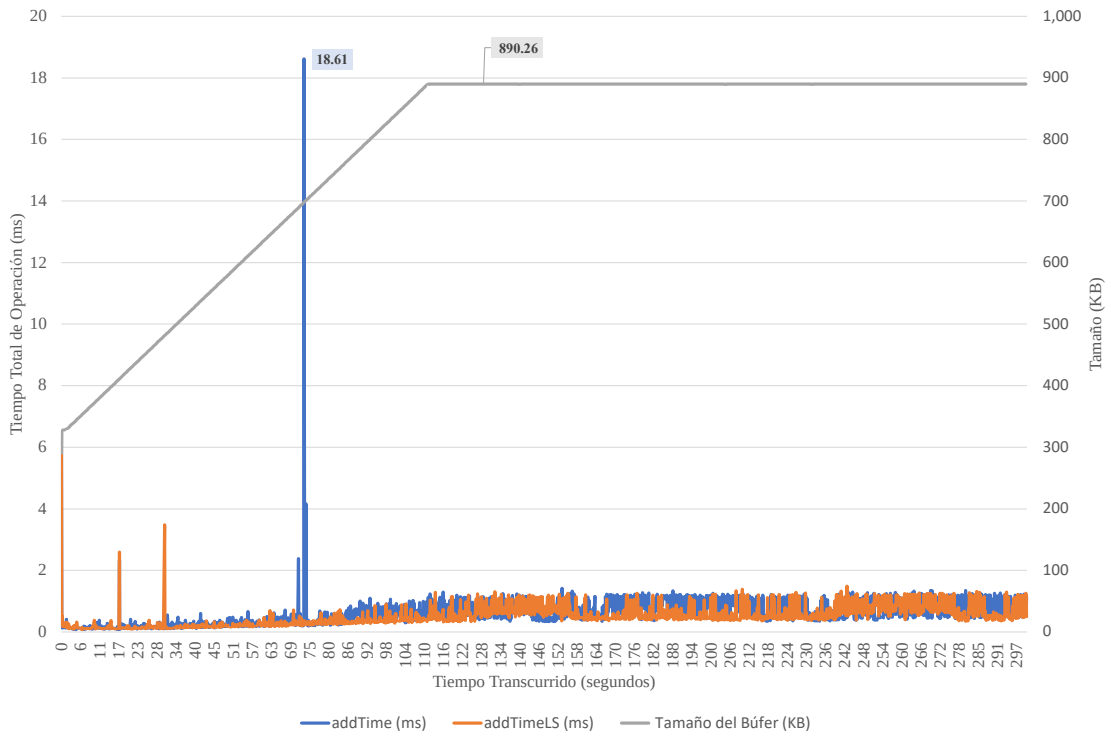

Figura 44 Evolución del tiempo total de operación y tamaño del búfer durante 5 minutos (Simulación 1 y 2)

La Figura 44 muestra la evolución de la operación “add” (es decir, agregar medidas al búfer) conjuntamente con el tamaño del búfer cuando el descarte selectivo está activado (es decir, simulación 2 descrito con la línea en color naranja) y desactivado (es decir, simulación 1 descrito con la línea en color azul). El tiempo de operación es cuantificado y expuesto utilizando el eje ordenado izquierdo, mientras que el tamaño del búfer emplea el eje ordenado derecho. Los valores atípicos son consecuencia del recolector de basura de JAVA. Independientemente de ello, es posible apreciar que la operación “add” tuvo un comportamiento similar aunque algunas diferencias pueden destacarse mejor mediante el gráfico boxplot de la Figura 45.

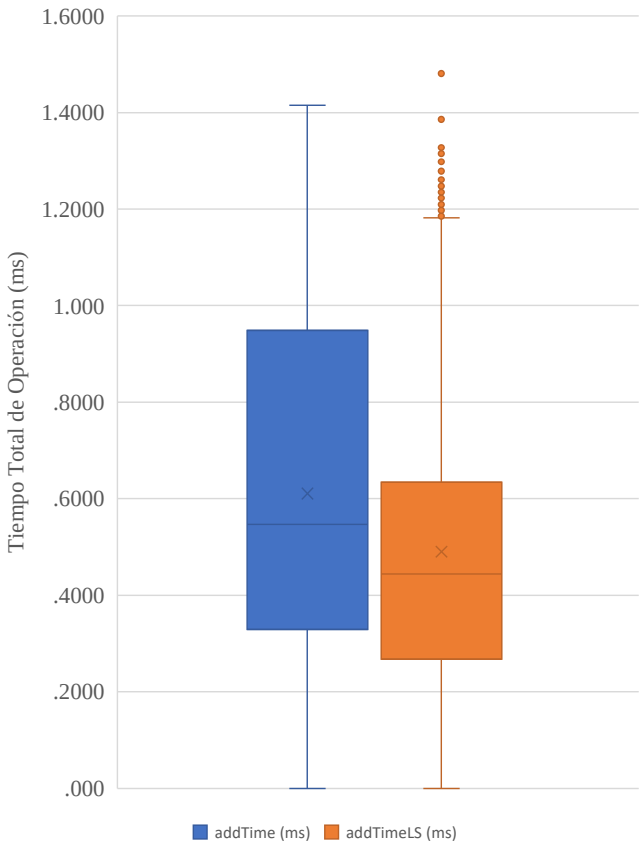

Figura 45 Boxplot para la operación "add" con descarte selectivo activado (addTimeLS) y desactivado (addTime)

Cuando el descarte selectivo está activado (caja naranja en Figura 45), los tiempos de la operación son sutilmente mejores, logrando una media aritmética de 0,49 ms y mediana de 0,44 ms con el descarte activo versus una media de 0,62 ms y una mediana 0,54 ms cuando está inactivo. Adicionalmente, con el descarte activo se logra una regulación del rango de variación para el tiempo de operación (es decir, el tamaño de la caja del boxplot), lo que es consistente con sus respectivas desviaciones de 0,31 ms para el descarte activo contra 0,51 ms cuando está inactivo. De este modo, el empleo de descarte selectivo decrementaría el tiempo de operación individual en 0,1 ms considerando la mediana, lo que representa una optimización de 22,72% (es decir, [[0.54/0.44]-1]*100).

Retomando la Figura 44, los valores atípicos podrían introducir limitaciones temporales a la aplicabilidad del descarte selectivo en conjunto con el búfer de datos. Es decir, las estimaciones incrementales serían posibles sí y solo sí existe 18,61 ms entre arribo de datos en un escenario pesimista. En un escenario optimista, el mejor tiempo de operación representó solo 0,08 ms. Es importante destacar que el tamaño máximo del búfer para las 10 métricas con 1000 slots (es decir, 10.000 medidas) requeriría 890,26 Kb para almacenar el búfer junto con las respectivas estimaciones.

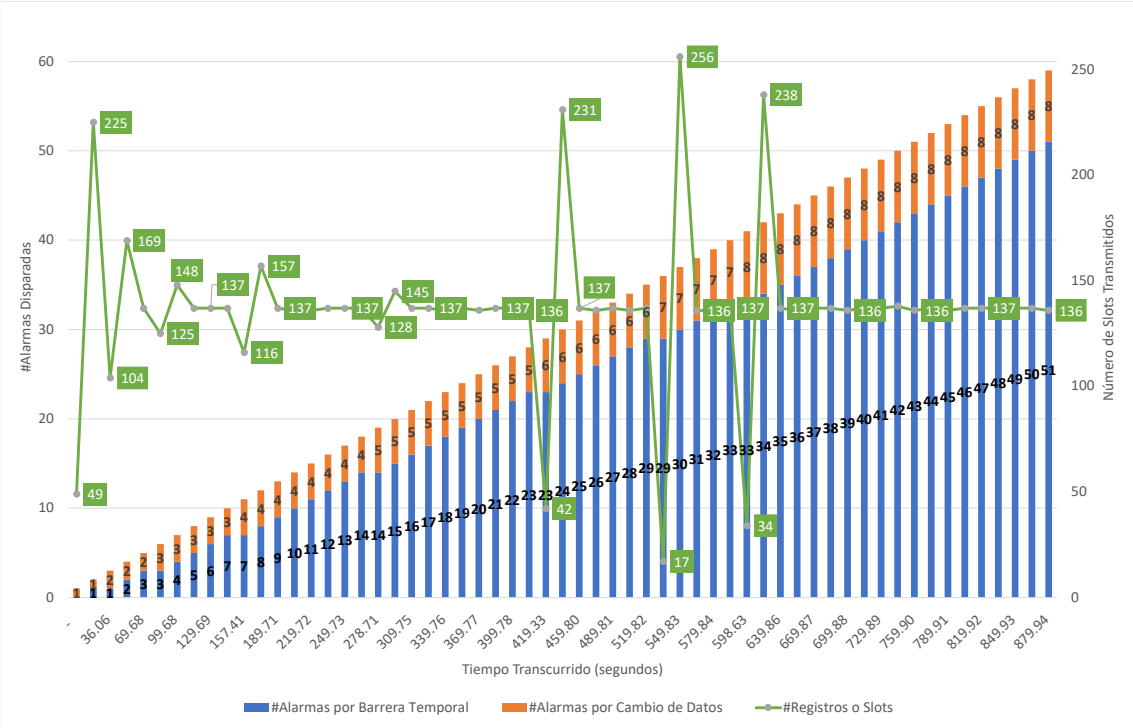

Figura 46 Evolución de las Alarmas Disparadas y Transmisiones de Datos durante 15 minutos (Simulación 3)

La Figura 46 describe el comportamiento descrito en la simulación 3 de acuerdo al número de alarmas disparadas y el volumen de transmisión de datos durante 15 minutos con el descarte selectivo desactivado. El eje ordenado izquierdo indica el número de alarmas disparadas, mientras que el eje ordenado derecho refiere al número de slots o registros transmitidos. Tenga en cuenta que cada slot o registro contiene un vector de datos completo para todas las métricas involucradas. Las alarmas por barreras temporal se indican con barras azul, mientras que aquellas relacionadas con cambios de datos se indican con barras naranjas. El número en la barra describe un acumulador de alarmas disparadas por tipo. Por ejemplo, el 51 de la última barra se asocia con la cantidad acumulada de alarmas por barrera temporal disparadas hasta ese momento, mientras que 8 es el análogo para las alarmas acumuladas por cambio en los datos. De este modo, De las 59 alarmas disparadas al final de los 15 minutos, 86,44% corresponden a barreras temporales y 13,56% a cambios en los datos.

Independientemente del origen de la alarma que genera la transmisión, un promedio de 136 slots se informa en cada operación lo que representa 1360 medidas (10 métricas por 136 medidas en cada slot). El volumen de datos mínimo informado fue 17 slots (170 medidas) consumiendo 14,26 kB, mientras que el volumen máximo fue 256 slots (2560 medidas) consumiendo 117,88 kB.

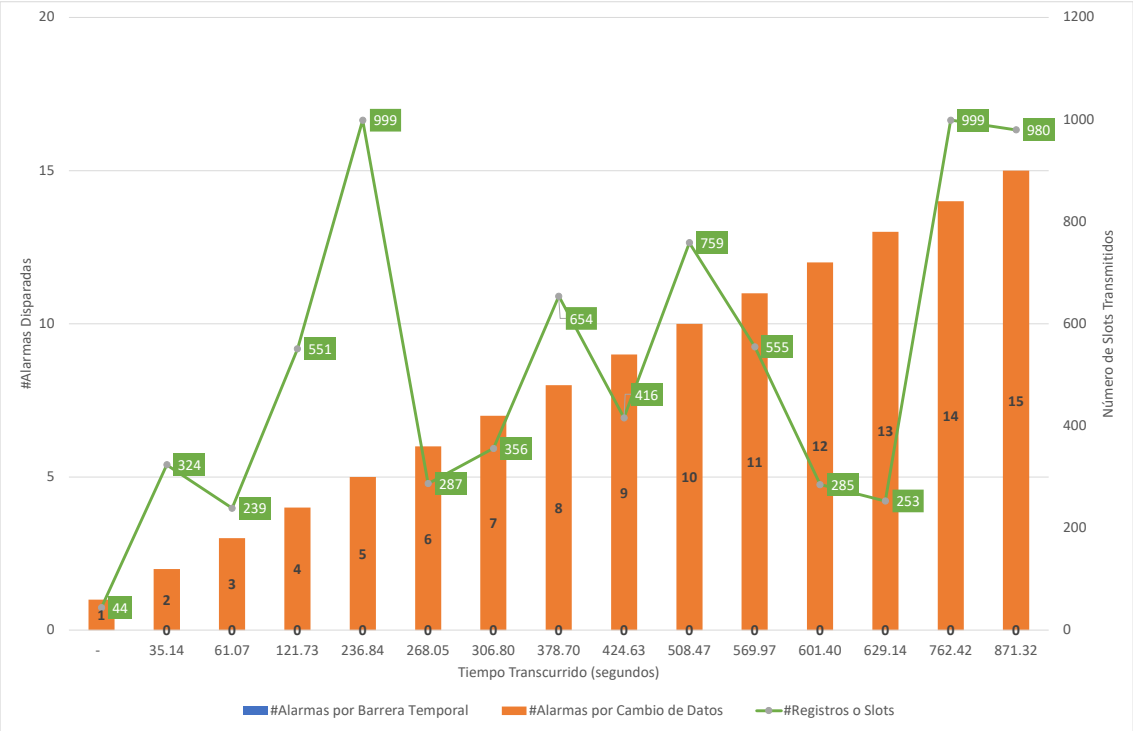

Figura 47 Evolución de las Alarmas Disparadas y Transmisión de Datos durante 15 minutos con descarte activado (Simulación 4)

La Figura 47 describe el comportamiento expuesto durante 15 minutos de simulación con descarte activo. El eje ordenado izquierdo indica el número de alarmas disparadas mientras que el eje ordenado derecho representa los slots transmitidos. Los resultados son prometedores dado que solo 15 transmisiones se requirieron versus las 59 de la simulación anterior. Esto evitaría 74,57% de las transmisiones de datos, consumiendo solo los recursos asociados con un 25,43% de las transmisiones originales. Además, el volumen informado de datos cambió. Cuando la técnica está activa, el número máximo de datos transmitidos ascendió a 999 slots (9990 medidas) consumiendo 440,05 kB, mientras que el número mínimo fue 44 slots (440 medidas) consumiendo 25,96 kB. Todas las alarmas en esta simulación se asocian con cambios en los datos detectados a través del filtro en línea basado en las medidas retenidas de acuerdo con las puntuaciones Z y ponderaciones asociadas.

Sin embargo, la Figura 48 indica el comportamiento expuesto para 15 minutos de simulación con descarte y barreras temporales inactivas. De este modo, las transmisiones de datos (sin ningún descarte) ocurren a partir de las alarmas de cambio de datos. El eje ordenado derecho indica la cantidad de alarmas disparadas, mientras que el derecho señala la cantidad de slots de datos transmitidos. Comparando la presente figura respecto de la Figura 47 es posible observar, que incluso cuando el número de transmisión de datos es la misma, la técnica de descarte (Figura 47) incorpora un retardo en la transmisión. Es decir, si se lee desde la segunda transmisión en adelante en ambas figuras, puede observarse que la segunda transmisión en con descarte ocurre

a los 35,14 segundos (Ver Figura 47) mientras que sin descarte ocurre a los 25,65 segundos (Ver Figura 48). Esto se sostiene en ambas figuras para toda la simulación.

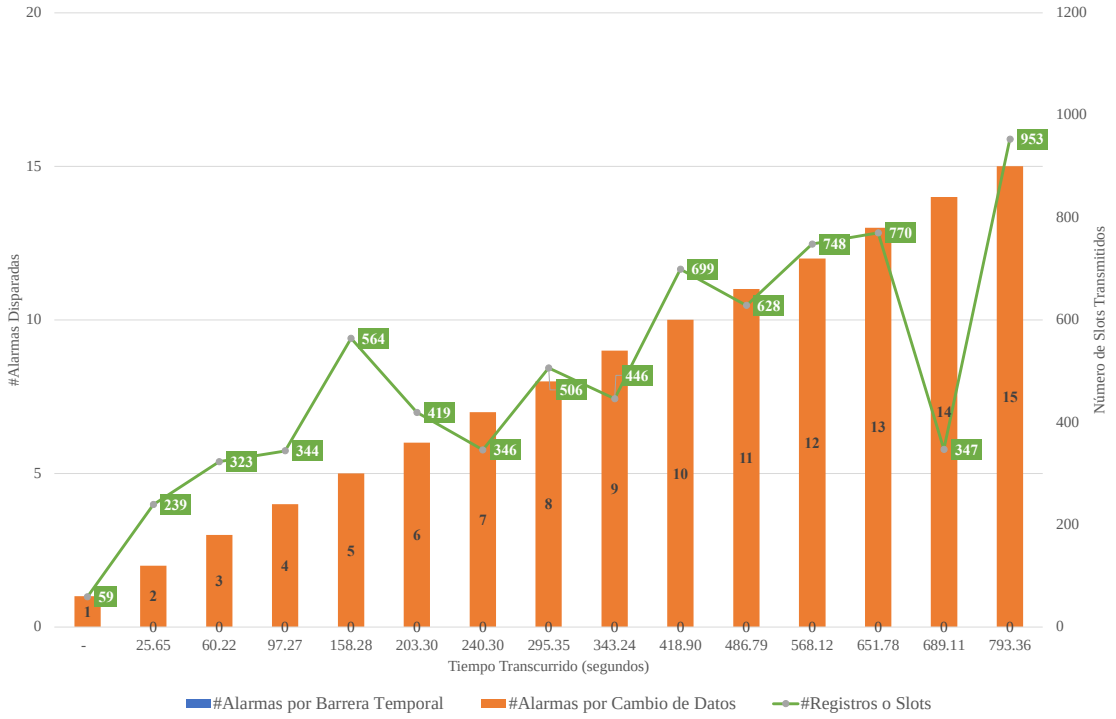

Figura 48 Evolución de las Alarmas Disparadas y Transmisión de Datos durante 15 minutos con descarte y barreras temporales inactivas (Simulación 5)

Por esa razón sería probable que las transmisiones de datos sin descarte fueren mayores para periodos de tiempo más grandes. Sin embargo, este resultado es interesante porque permitiría omitir el descarte de medidas y utilizar transmisiones basadas en cambios en los datos para colaborar con la reunión central de medidas en APbMM ante potenciales limitaciones en la tasa de procesamiento.

6.3 Registro de Integridad basado en Merkle Tree

Esta sección describe la relación entre el flujo de medidas, el control de integridad, y el árbol de Merkle. Adicionalmente, se proveen simulaciones de referencia como guía para la aplicabilidad de la propuesta.

6.3.1 Flujo de Medidas y Árbol de Merkle

Un árbol de Merkle es un árbol binario donde cada hoja contiene una huella (hash) de los datos (pero no los datos en sí), mientras que cada nodo intermedio hasta la raíz contiene un nuevo hash a partir de sus nodos hijos [154].

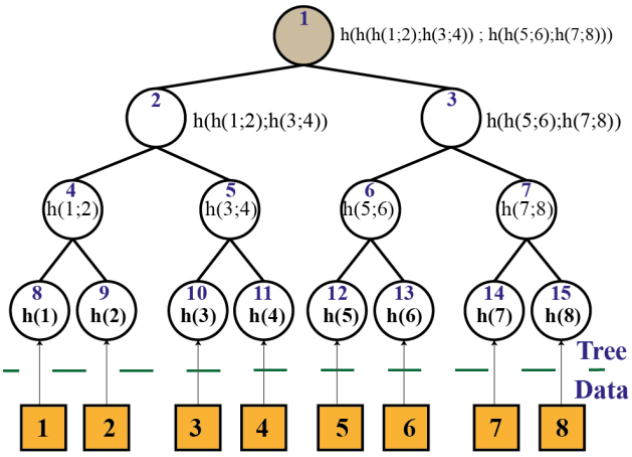

Figura 49 Un típico árbol de Merkle

Como se puede apreciar en la Figura 49, la datos podría ser un conjunto de transacciones o un gran archivo separado en trozos o partes. La estructura de dato no contiene el dato en sí mismo sino su hash asociado. De este modo, cada hoja del árbol contiene el hash asociado con su trozo de datos, mientras que su padre contendrá el hash calculado a partir del hash de sus hijos. Por ejemplo, el nodo indicado con un número 4 en la anterior figura posee su hash calculado a partir de los nodos 8 y 9, y de este modo sucesivamente hasta alcanzar la raíz. El esquema de Merkle permite verificar la integridad del dato en un entorno distribuido a través de unas pocas operaciones basadas en el cálculo de hash, mientras que el resto de los datos no son requeridos. Por ejemplo, suponga que un gran archivo de datos es trozado en 8 partes y se distribuyen entre diferentes nodos de un cluster. Por alguna razón, se necesita verificar la integridad del trozo número 4 (Ver nodo 11 en Figura 49). Para ese propósito, el nodo correspondiente utilizará los hashes relacionado al nodo 10 (es decir, h(3)), 4 (es decir, h(1;2)), y 3 (es decir, h(h(5;6);h(7;8))). De este modo, mediante la comparación de h(h((h(3); h(4));h(1;2));3) con el hash de la raíz del árbol, se podría saber si el trozo de datos es válido o no en base a si emparejan o no sus huellas.

Como se introdujo en la sección 5.2.1, el formato Brief contiene una huella MD5 como encabezado del mensaje que permite verificar la integridad del mensaje (considerando las medidas y la definición del proyecto), pero nada dice sobre los mensajes previos (Ver Figura 25). En otras palabras, el adaptador de medición y la recolección de datos en PAbMM tienen un medio a través del cual es posible saber si un mensaje ha sido modificado o no aunque está limitado a un único mensaje.

Así, dado que el mensaje Brief contiene una huella MD5 como encabezado de mensaje, es posible utilizar esta como parte de un árbol de Merkle para mantener un seguimiento de un cierto número de transacciones pasadas, evitando tener que recalcular la huella del mensaje. Esto es útil tanto para el adaptado de mediciones para mantener un registro local de las ventanas de datos enviadas, como la función de recolección de datos de PAbMM para mantener un seguimiento de los datos recibidos desde estos.

De este modo, se introduce el empleo del árbol de Merkle para implementar un registro de longitud fija para verificar la integridad de las últimas 2^n transacciones, donde "n" representa la profundidad del árbol. Así, la profundidad del árbol depende del número de transacciones a mantener dentro del registro.

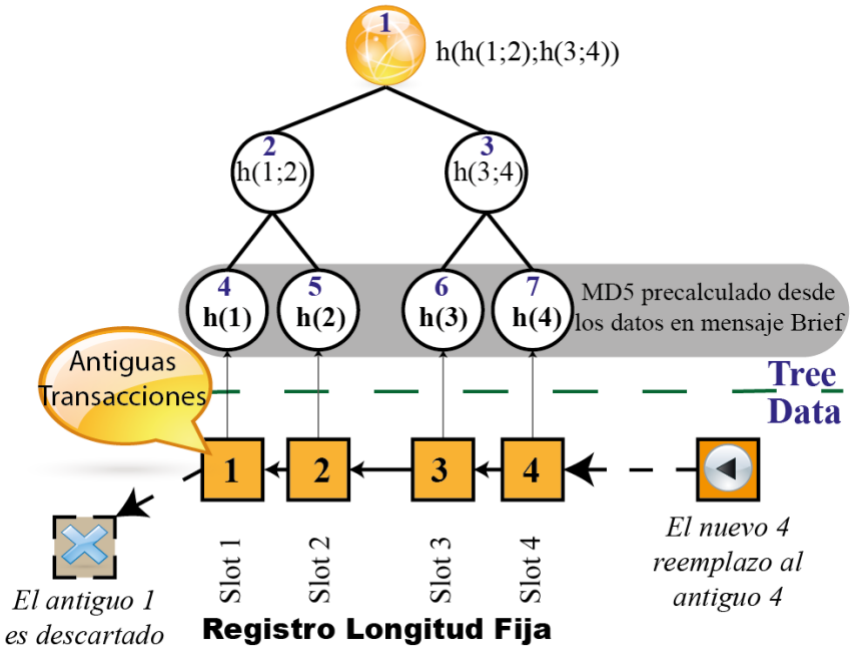

Figura 50 Un árbol de Merkle Orientado a Soportar un Registro de Longitud Fija para Verificación de Integridad

La Figura 50 describe la representación de un registro de integridad con una capacidad para almacenar las últimas cuatro transacciones. Las transacciones antiguas se localizan en el extremo izquierdo de las hojas, mientras que las transacciones recientes lo hacen en el extremo derecho. Cuando una transacción es incorporada, la transacción más antigua es descartada (es decir, el nodo 1 en la Figura 50) para dar su lugar al nodo que le sigue en antigüedad (nodo 2), quien se torna en el nuevo nodo más antiguo. Del mismo modo, el nodo 3 se torna en el nuevo nodo 2, el nodo 4 se torna en el nuevo nodo 3, y finalmente, la nueva transacción ocupará el slot 4. En este registro, las hojas no necesitan calcular su huella debido a que el MD5 viene dentro del encabezado del mensaje Brief. Sin embargo, los nodos intermedios hasta la raíz necesitan recalcular las huellas.

Este registro simplifica la verificación de integridad entre los adaptadores de medición (AM) y la función de recolección (FC) de PAbMM, permitiendo contrastar la historia reciente basado en la huella MD5 de la raíz del árbol. Es decir, AM y FC necesitan solo comparar las huellas MD5 para saber si existe integridad. Adicionalmente, es posible realizar verificaciones parciales a través de los nodos intermedios. Por ejemplo, la huella de las últimas dos transacciones se almacena en el nodo 3 (Ver Figura 50), y análogamente, contrastando el MD5 del nodo 3 con su respectivo de FC es posible verificar su integridad.

Dado que AM es un componente localizado en dispositivos con recursos limitados, el volumen de transacciones a monitorear dependerá de las capacidades del dispositivo. Incluso, el volumen de datos a transmitir es típicamente alto en Internet de las Cosas, razón por la cual el registro es inicializado en el arranque y mantenido en memoria hasta que el dispositivo se apaga. Así, durante el booteo se crea un árbol binario denso y se deja listo para ser empleado con las transacciones basadas en mensajes Brief.

La implementación del árbol de Merkle basada en un arreglo unidimensional se encuentra disponible en el repositorio mair de GitHub bajo los términos de la licencia Apache 2.0. La librería se denomina mair como acrónimo para la expresión en ingles de Registro de Integridad del Adaptador de Medición (en inglés, *Measurement Adapter Integrity Record*). Allí se incorpora tanto la perspectiva local al AM como la global para articular la FC con un conjunto de AM.

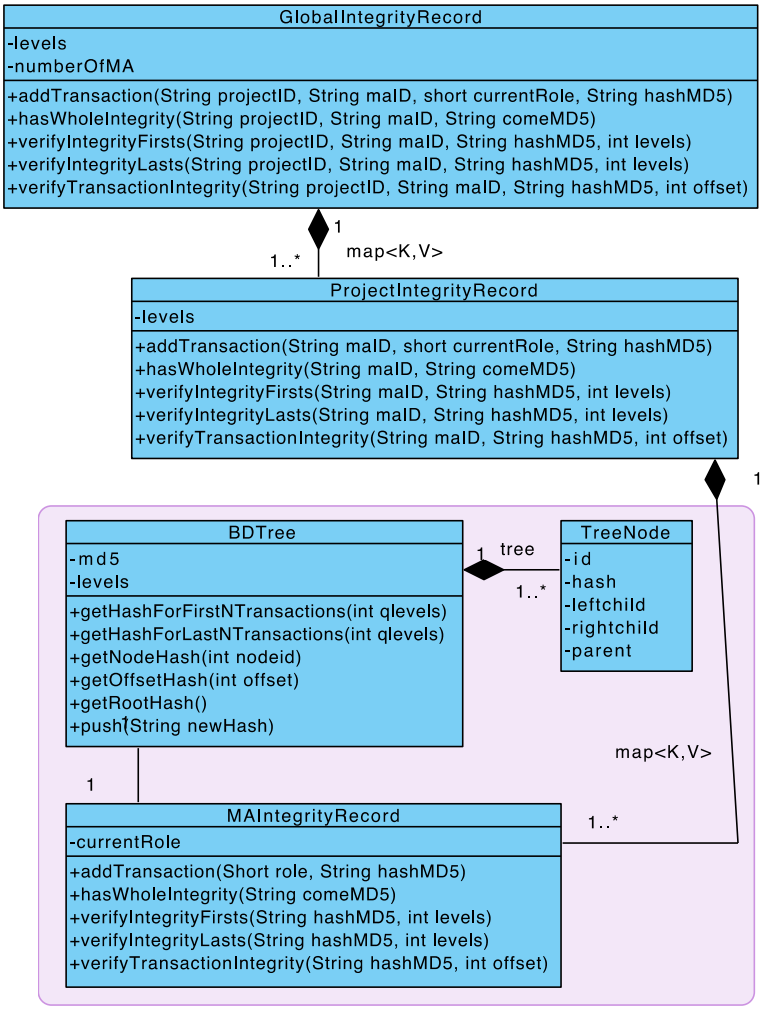

Figura 51 Clases Principales Relacionadas a la Librería MAIR

La Figura 51 describe las principales clases asociadas con la librería MAIR, dentro de las que se encuentran a) *TreeNode*: representa la estructura de datos mínima donde la huella es almacenada; b) *BDTree*: implementa la lógica del árbol de Merkle como un arreglo unidimensional a partir de un conjunto de instancias de TreeNode. El atributo

md5 es una instancia de la clase *java.security.MessageDigest*, mientras que el atributo *levels* se establece como valor por defecto para la profundidad del árbol; c) *MAIntegrityRecord*: implementa el registro de integridad en el adaptador de medición. Se emplea una instancia de la clase *BDTree* para mantener un seguimiento de las transacciones informadas; d) *ProjectIntegrityRecord*: Implementa el registro de integridad por proyecto. Dado que un proyecto puede contar con un conjunto de AM asociados, una tabla hash utiliza el ID del AM para acceder a su árbol de Merkle asociado; e) *GlobalIntegrityRecord*: Implementa el control de integridad a nivel global (múltiples proyectos). Por ellos, se emplea una tabla hash utilizando el ID de proyecto como clave de acceso a su registro de integridad.

En particular, la clase *MAIntegrityRecord* contiene un atributo *currentRole* que representa el último rol conocido para el adaptador que es representado. En esta clase, las principales responsabilidades pueden sintetizarse como sigue:

- **addTransaction**: Incorpora un nuevo hash al registro de transacciones, provocando un desplazamiento a la izquierda de los antiguos registros para generar el espacio necesario (descarta el más antiguo si fuere necesario). Actualiza las huellas encadenadas en forma jerárquica a lo largo del árbol hasta alcanzar la raíz. Finalmente, el rol actual del AM es actualizado (es decir, gateway, blocked, cooperative, o data collector).

- **hasWholeIntegrity**: El método compara la integridad global del registro. Por esa razón, compara la huella que se recibe como parámetro con la localizada en la raíz del árbol para saber si emparejan o no. Cuando las huellas emparejan, se dice que ambos registros tienen integridad. De otro modo, existe una diferencia independientemente del origen de la misma.

- **verifyIntegrityFirsts**: El método permite comparar la integridad de un subconjunto de transacciones comenzando desde el más antiguo (es decir, de izquierda a derecha en la Figura 50). Se compara la huella relacionada con los nodos intermedios que contienen los 2^{levels} transacciones con la huella indicada como un parámetro para saber si empareja o no. Cuando las huellas comparadas coinciden, se dice que el conjunto de 2^{levels} transacciones tiene integridad entre registros. De otro modo, alguna transacción entre las comparadas contiene una diferencia, aunque no focaliza en saber cuál es su origen.

- **verifyIntegrityLasts**: El método es similar al anterior, solo que el presente se focaliza en el subconjunto de transacciones comenzando desde la más reciente (es decir, de derecha a izquierda en el árbol de la Figura 50). Así, se compara la huella relacionada con el nodo intermedio que contiene las últimas 2^{levels} transacciones contra la huella indicada como parámetro para saber si

emparejan. Cuando las huellas comparadas emparejan se dice que los registros poseen integridad. En caso contrario, los registros no poseen integridad sin interesar en este punto el origen de la diferencia.

- **verifyTransactionIntegrity**: Este método contrasta dos huellas de diferentes registros para una misma transacción. Así, cuando las huellas coinciden, se dice que la transacción tiene integridad y corresponde al mismo contenido en ambos registros.

El rectángulo en la Figura 51 indica la funcionalidad que reside en el adaptador de medición. El resto de las clases no contempladas corresponden al registro de integridad global utilizado en la PAbMM. Las funcionalidades de los métodos son las mismas, cambian los parámetros debido al nivel de granularidad que se gestiona.

De este modo, el adaptador de medición puede mantener un seguimiento de cada ventana de datos transmitida hasta un tamaño dado. El tamaño del registro de integridad dependerá de las capacidades del hardware en el que se ejecuta. Así, AM se torna en un nuevo lugar donde la función de recolección puede verificar la integridad de las ventanas de datos recibida a través de un simple contraste de huellas. Adicionalmente, es útil para ser utilizado como verificación de segundo factor, dado que independientemente la ventana de datos que se reciba en PAbMM, éste podría verificar contra AM la huella de su raíz para saber si el mensaje fue modificado.

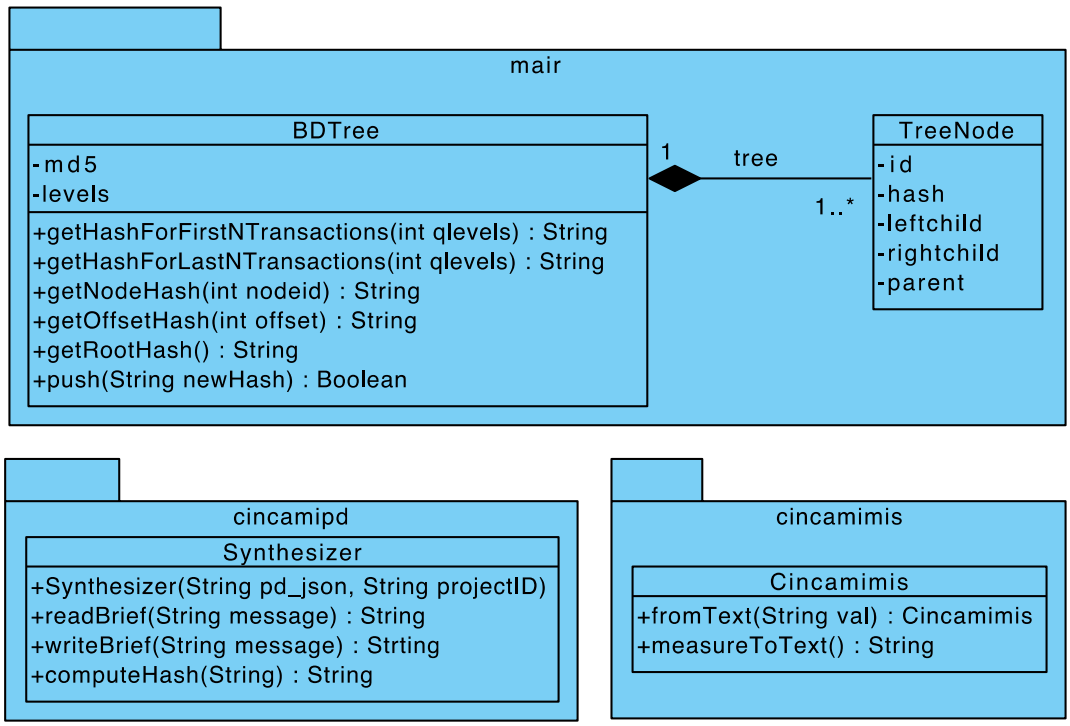

Figura 52 Paquetes Necesarios para Implementar la Relación entre el Mensaje de Datos Brief y el Registro de Integridad de PAbMM

La Figura 52 sintetiza las librerías utilizadas por PAbMM para implementar el mensaje de datos Brief guiado por metadatos conjuntamente con el registro de integridad. En este sentido, los puntos en común entre las librerías residen en la definición del proyecto (es decir, la librería *cincamipd*) el cual establece los conceptos a monitorear y el modo en que se cuantifica cada atributo o propiedad de contexto. La librería *cincamimis* es reposnable por generar y leer el mensaje de intercambio (sea como XML, JSON, o Brief). La clase *Synthesizer* implementa el cómputo de huellas MD5 desde el contenido del mensaje para producir el MD5 articulando la definición del proyecto y mensaje con las medidas.

6.3.2 Patrones de Referencia

Para proveer una perspectiva cuantitativa respecto de la propuesta, se llevaron a cabo dos simulaciones ejecutadas en una MacBook Pro con a macOS Catalina 10.15.4, 16GB de RAM LPDDR3 2133Mhz.

La primera simulación analizó el tiempo y tamaño involucrado en la creación del registro de integridad global. Para ese propósito, se definen diferentes ID de proyectos de medición en una lista, mientras que se indican a tareas Runnable en Java. La variación de parámetros se asocia con el número de proyectos simultáneos, número de adaptadores por proyecto, y volumen de transacciones por adaptador. Al inicio, cada hilo procesa la creación para una configuración dada (combinación de proyectos simultáneos, número de adaptadores, y número de transacciones por adaptador) midiendo el tiempo requerido y su tamaño.

La segunda simulación analizó los tiempos de operación individuales del registro de integridad. Así, el registro de integridad global se crea y llena con transacciones a lo largo de 20 minutos. Se analizaron las operaciones a) addTransaction, b) hasWholeIntegrity, c) verifyIntegrityFirsts, d) verifyIntegrityLasts, y e) verifyTransactionIntegrity.

La Tabla 30 describe alguna de las principales combinaciones para estimar el tiempo total requerido para inicializar el registro global de integridad junto con su tamaño requerido. Se indica el número de proyectos simultáneos bajo monitoreo en el registro de integridad, con una cierta cantidad de adaptadores de medición (AM) que informan medidas en forma concurrente.

Con el objetivo de limitar el tamaño de la tabla, se exponen los casos para 1024 transacciones por AM. Se supone que cada transacción contiene 200 medidas (1 por cada segundo) y ocurre cada 200 segundos (alrededor de 3 minutos). Así, las 1024 transacciones podrían almacenar alrededor de 3072 minutos (es decir, 1024 trx *3 min/trx=3072, lo que sería alrededor de 51,2 horas o 2,13 días) que representa una capacidad interesante. De todos modos, el tamaño del registro es un parámetro y puede ser ajustado acorde a cada situación.

Tabla 30 Registro de Integridad: Tiempo de Creación Consumido y Tamaño Requerido

#Proyectos	#AM con 1024 Transacciones	Tiempo Total (ms)	Tamaño Total (kB)	Tamaño por Transacción
1	1	4,85	196,00	0,1914
1	50	236,00	7369,00	0,1431
1	100	238,00	14688,00	0,1434
1	200	1032,00	29327,00	0,1432
5	50	1726,00	36647,00	0,1432
5	100	3007,00	73243,00	0,1430
5	200	6078,00	146436,00	0,1430
10	50	3563,00	73244,00	0,1430
10	100	5252,00	146435,00	0,1430
10	200	22605,00	292822,00	0,1429
15	100	17599,00	219630,00	0,1429

La primera línea de la Tabla 30 representa la situación donde cada adaptador de medición mantiene un registro local con una capacidad de almacenamiento para 1024 transacciones. Sería necesario entonces 196 kB para almacenar las mismas y consumiría alrededor de 4,85 ms su inicialización. Tanto el tiempo de creación como el tamaño requerido hacen del registro un elemento interesante para emplear a nivel de AM.

Cuando se requieren monitorear 10 proyectos en simultáneo recibiendo medidas desde 200 adaptadores (cada uno con capacidad de 1.024 transacciones), el tiempo de inicialización requerido es 22.605 ms (alrededor de 22,6 segundos) y 292.822 kB (285,95 MB). En esta configuración, el registro podría almacenar hasta 2.048.000 transacciones (es decir, 10 * 200 * 1024).

Adicionalmente, la tabla expone el tamaño por transacción como el cociente entre el tamaño total requerido y el número total de transacciones soportadas. Un razonamiento análogo podría implementarse con el tiempo requerido por transacción para crear el registro. De este modo, a partir de los resultados, puede indicarse que sería necesario 0,143 kB y [0,005; 0,019] ms por transacción para crear el registro.

Ecuación 32 Estimación del Tamaño Requerido para el Registro de Integridad

$$Total_{size} = nproj * nma * ntrx * 0.143$$

La Ecuación 32 representa una aproximación al cálculo del tamaño total requerido por el registro de integridad para una configuración dado, donde se tiene que:

- **nproj**: Representa el número total de proyectos a monitorear en el registro.

- **nma**: Indica el número de adaptadores por cada proyecto a considerar en el registro.

- **ntrx**: Refiere al número de transacciones máximas a almacenar por cada adaptador.

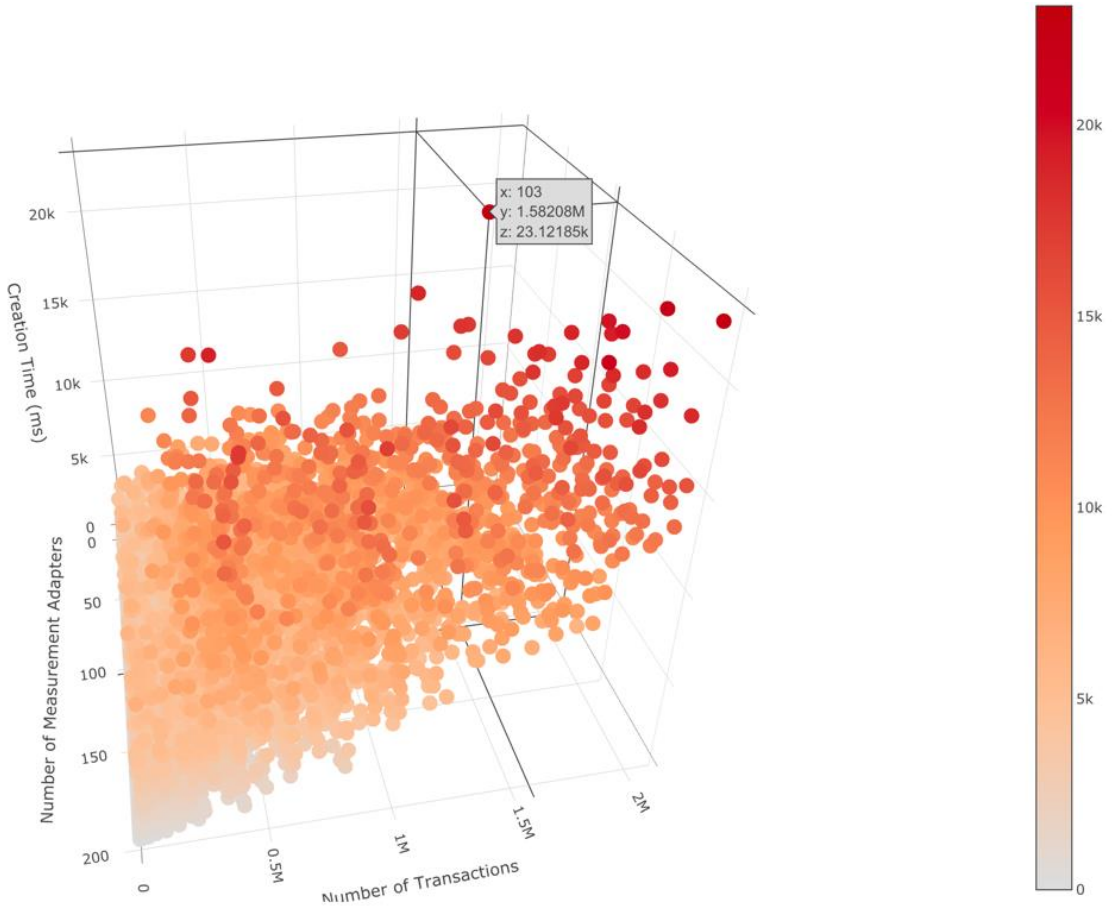

Figura 53 Gráfica de Dispersión para la creación del registro. Perspectiva Superior Tiempo de Creación y Número de Transacciones.

Como puede observarse para la nube de puntos de la Figura 53, pareciera que existe una relación aparentemente lineal entre el tiempo de creación y el número de transacciones.

En principio, pareciera que el número de adaptadores de medición no es relevante en forma directa, sino que lo es a través del volumen de transacciones que aporta.

Es decir, el tiempo de creación no se asociaría a mayor o menor número de adaptadores de mediciones, sino que lo que impacta es el volumen de transacciones para las que debe construir las estructuras de dato de soporte. Las estructuras de soporte intermedia para los adaptadores no aportarían demasiado al tiempo total.

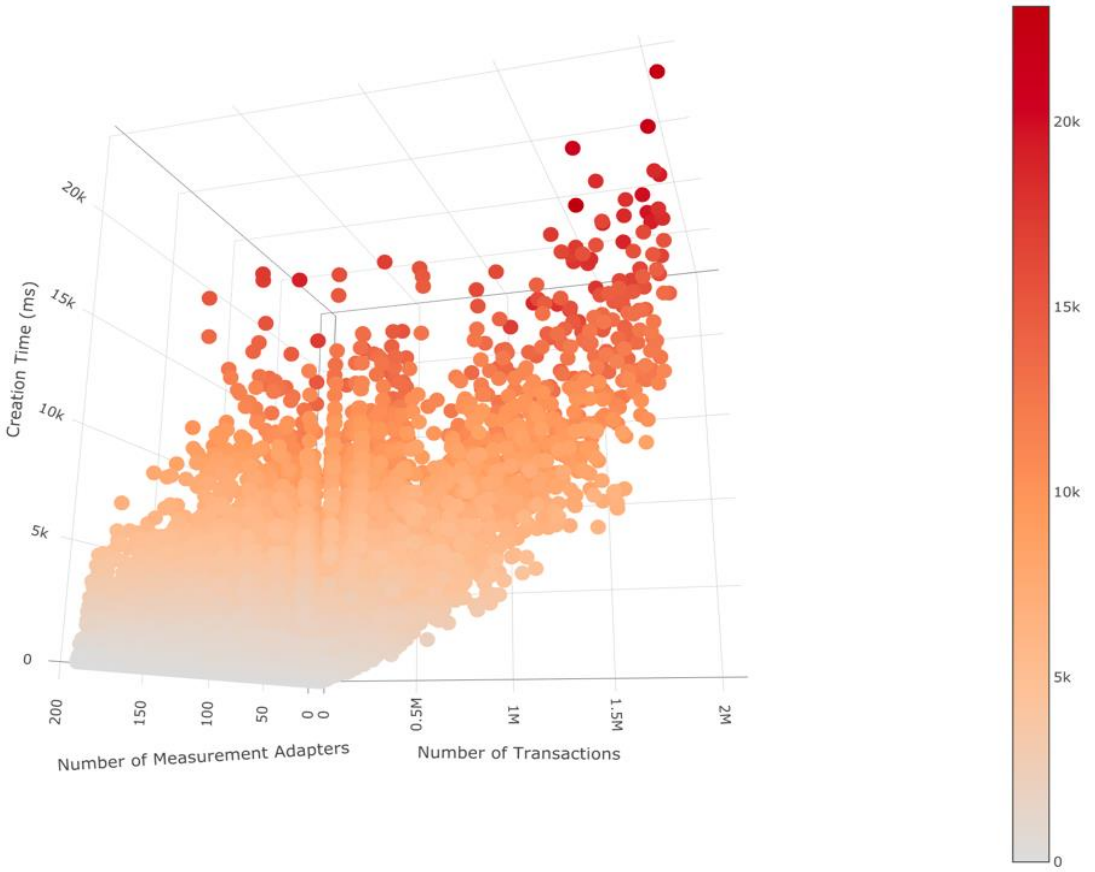

Figura 54 Gráfica de Dispersión para la creación del registro. Perspectiva Inferior Tiempo de Creación y Número de Transacciones.

La Figura 54 expone una vista inferior de la nube de puntos, donde puede observarse la inclinación de la nube de puntos hasta alcanzar las 2 millones de transacciones.

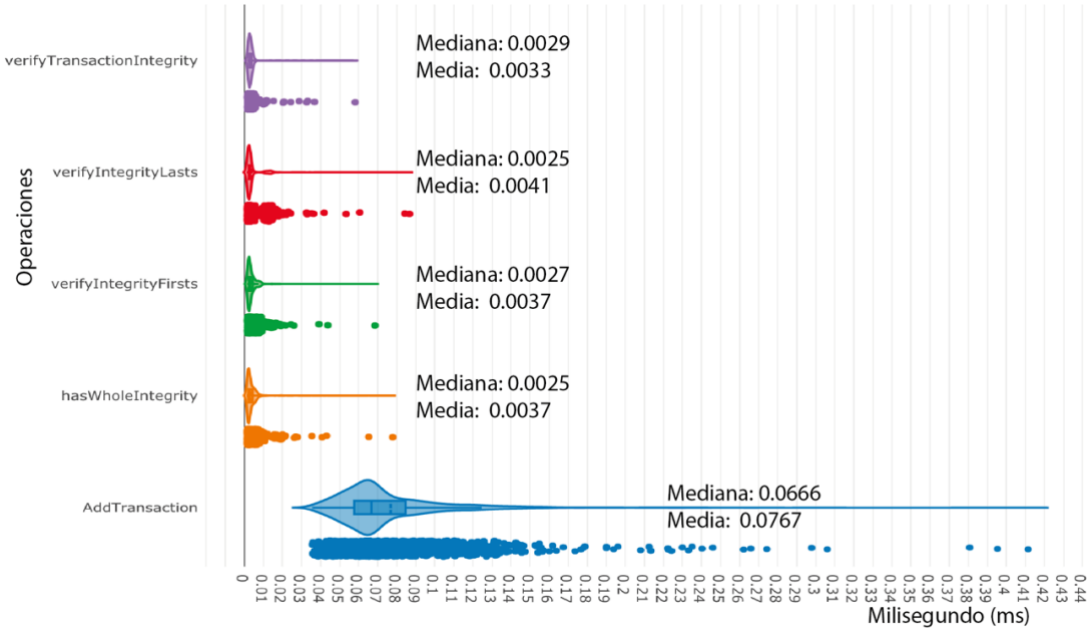

Figura 55 Tiempo Consumido por las Operaciones Individuales en el Registro de Integridad a lo largo de 20 minutos

La Figura 55 sintetiza las gráficas de violín para los tiempos de operación medidos a lo largo de 20 minutos en la segunda simulación. Como se puede apreciar, la operación más costosa desde el punto de vista temporal fue *addTransaction* con una mediana de 0,0666 ms y una media de 0,0767 ms. Como puede apreciarse, la diferencia entre media y mediana se debe a la presencia de valores atípicos superiores que empujan hacia arriba el valor de la media.

Por otro lado, el resto de las operaciones consumieron entre 0,0025 ms y 0,0029 ms de acuerdo con sus medianas. También presentan medias superiores a la mediana debido a la presencia de valores atípicos superiores. Estos se vinculan con el recolector de basura de Java.

Por un lado, el comportamiento es lógico con lo que se esperaría dado que la operación que más consume (es decir, *addTransaction*) implicaría un eventual movimiento de las transacciones y un recálculo de las huellas del árbol. Por otro lado, las restantes operaciones se limitan al cálculo de una huella y su comparación, sin alterar el árbol de Merkle de ningún modo.

6.4 Registro Distribuido de Adaptador de Mediciones basado en Blockchain

Esta sección incorpora la tecnología de cadena de bloques para soportar el registro unificado de nodos de forma distribuida e independiente de las capas en la nueve de PAbMM. Ello dota de autonomía a los nodos (adaptadores o pasarelas) permitiendo acortar la ruta de acceso a recomendaciones o definiciones de proyecto y calcular similitudes basado en la distancia compuesta localmente al nodo (en lugar de hacerlo solo en forma centralizada). La primera parte describe la organización del registro distribuido basado en cadena de bloques. La segunda sección describe su perspectiva comportamental. La tercera sección provee resultados de consumo de memoria y tiempos de inicialización como patrón de referencia basado en dos simulaciones discretas.

6.4.1 Registro Distribuido basado en Blockchain

Como se mencionó en el capítulo 5, en el campo de la recolección de medidas se tienen dos tipos de perfiles: el adaptador de medición (o puente semántico) y la pasarela cada uno de los cuales posee una configuración de hardware distintiva. Por un lado, la pasarela actúa de concentrador de datos minimizando la dependencia de cada adaptador de medición de la nube. Por otro lado, el adaptador opera con los sensores directamente, recolectando las medidas, procesándolas y generando los mensajes de intercambio (Ver Sección 5.2). La transmisión de datos podría ser realizada directa o indirectamente a partir del adaptador (Ver sección 5.2.3).

Los adaptadores de medición y las pasarelas requieren de un conocimiento completo de los elementos que participan en la recolección de datos. La Figura 56 describe una perspectiva de despliegue, donde tanto las pasarelas como los adaptadores de medición

requieren de un componente denominado *LocalResources* para obtener acceso e interactuar con la cadena de bloques (o Blockchain). Este componente permite acceder a la cadena de bloques actual, el registro de transacciones junto con el árbol de Merkle describiendo las transacciones enviadas desde el nodo local. Como una diferencia con el árbol de Merkle, el registro de transacciones (es decir, el local) contiene todas las transacciones enviadas, pero también aquellas recibidas para aprobación desde la red de recolección. A partir de dicha información, el componente provee detalles sobre los nodos actuales, la huella de la raíz actual al árbol de Merkle, puntuación de los nodos (adaptadores o pasarelas), evaluación sobre la existencia del nodo, localizador de recursos uniformes (*Uniform Resource Locator* -URL) para el nodo con mayor puntuación, última huella del contenido para el último bloque, consumo de memoria por nodo, información del nodo local, contenido firmable, y el modo de leer la cadena de bloques completa (es decir, requestedAccess).

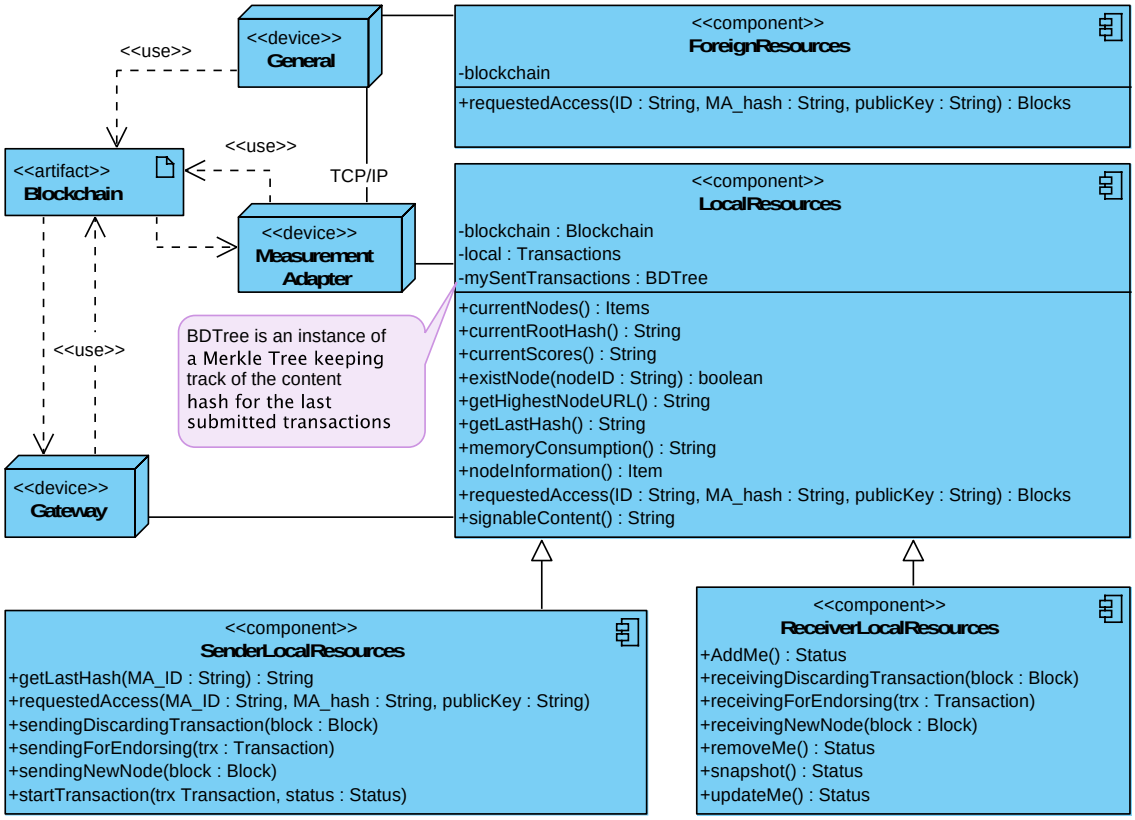

Figura 56 Una Perspectiva de Despliegue del Registro Distribuido basado en Blockchain

Por un lado, el componente *SederLocalResources* representa una especialización del componente *LocalResources* proveyendo servicios adicionales tales como el envío de la transacción para su aprobación, la posibilidad de descartar transacciones no aprobadas, la incorporación de nuevos bloques verificando las aprobaciones respectivas, y el inicio de una transacción (es decir, operaciones tales como Add, Remove, Snapshot, o Update). Por otro lado, el componente *ReceiverLocalResources* provee la lógica para

procesar cada tipo de solicitud (es decir, descartar y aprobar una transacción, o incorporar un nuevo nodo en la cadena de bloques).

Los adaptadores y pasarelas solo pueden modificar, incorporar, o borrar sus propios datos. Esto implica que nadie podría pretender iniciar una transacción en nombre de otro nodo. Así, todos los participantes tienen acceso a la última información provista por cada nodo para conocer, por ejemplo, si desea colaborar con transmisiones de medidas indirectas (por ejemplo, se encuentra en modo cooperative). Adicionalmente, es importante mencionar que la base de datos basada en blockchain es independiente de la arquitectura en la nube, proveyendo un registro unificado y autogestionado en el campo de monitoreo. Desde la arquitectura basada en la nube, la cuestión esencial es que la medida arribe y esto es esencial para la distancia compuesto a los efectos de estimar comportamientos de métricas. Sin embargo, desde la perspectiva del adaptador o la pasarela, el punto es transmitir las medidas del modo más inteligente y lo antes posible. De este modo, el registro distribuido provee cierto nivel de autonomía a cada pasarela o adaptador para decidir el modo de enviar sus medidas a la nube.

Adicionalmente, la anterior figura contempla el rol de dispositivo general (*General Device*) para acceder a la cadena de bloques en modo solo lectura y tomando un rol pasivo de observador. Es decir, este tipo de dispositivos podrían recuperar la cadena de bloques para conocer el estado actual de nodos que participan en la recolección de datos, pero no se les permite ejecutar ningún tipo de operación más que la descrita. Esto es útil cuando se requiere algún comportamiento de auditoría a implementarse sobre el registro distribuido y la transmisión de medidas.

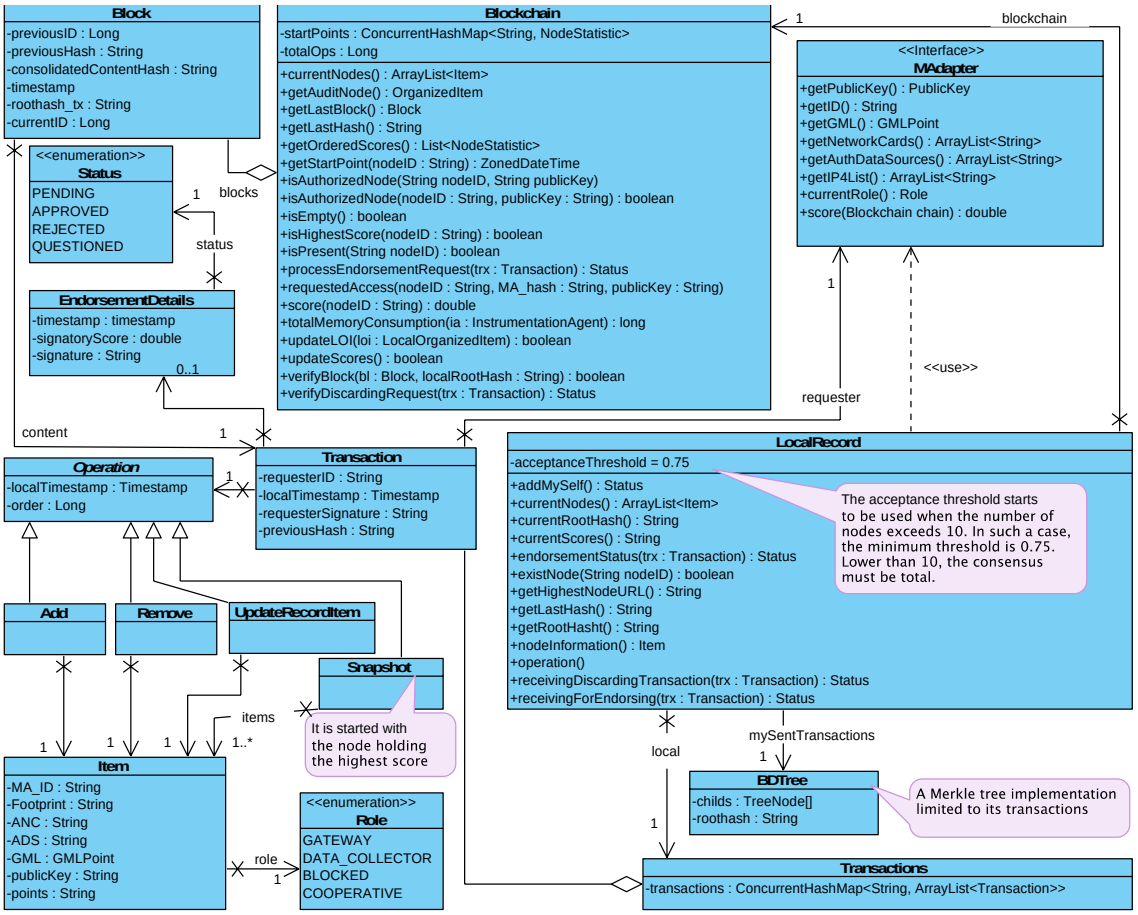

Figura 57 Principales Conceptos Relacionados con el Registro de Nodos Distribuidos basados en Blockchain

La Figura 57 describe los principales conceptos para implementar el registro de nodos distribuidos basado en Blockchain. Como su nombre lo indica, Blockchain se compone de un conjunto de bloques, donde cada uno tiene enlace al nodo previo y su huella, el contenido propio, la marca de tiempo cuando fue incorporado, y la huella de la raíz relacionada con el árbol de Merkle al momento en que se realizó la solicitud de incorporación (Ver clases *Block* y *Blockchain*). La clase *Blockchain* provee la funcionalidad asociada con la información obtenida desde este, por ejemplo, una vista materializada con el último estado conocido de todos los nodos (Las principales funcionalidades se sintetizaron en la Figura 56). Cada bloque tiene una transacción asociada (Ver la clase *Transaction*) donde el solicitante se identifica junto con el momento en que lo solicita, su firma, y la huella previa relacionada al contenido del bloque. Por cada transacción, una operación se asocia con la correspondiente información del ítem (Ver la clase *Operation*). La clase *Operation* describe la clase base desde donde heredan las clases Add, Remove, UpdateRecordItem, y Snapshot, describiendo las cuatro operaciones disponibles sobre la cadena de bloques. En otras palabras, un ítem puede agregar, remover, o actualizar su información, pero solo el nodo con mayor puntuación puede llevar adelante una snaphost (o instantánea).

La puntuación por nodo se calcula considerando el número de transacciones consolidadas, el tiempo desde el cual el nodo se encuentra activo, y el número de transacciones cuestionadas tal y como expone la Ecuación 33.

Ecuación 33 Cálculo de la Puntuación Por Nodo basado en Actividad en la Cadena de Bloques

$$score(node) = \left(\frac{\#Trx_{approved}}{\#Trx\_total} + \frac{Elapsed\_Time}{\#Total\_Time}\right)\Big/_{2}$$

Donde:

- **#Trx$_{approved}$**: Representa el número total de transacciones aprobadas para el nodo.

- **#Trx_total**: Indica el número total de transacciones ejecutadas sobre la cadena de bloques.

- **Elapsed_Time**: Contempla el tiempo total en el cual el nodo ha pertenecido a la cadena de bloques.

- **#Total_time**: Describe el tiempo total en el que la cadena de bloques ha estado activa.

El nodo con la puntuación (o por su término en inglés, *score*) más alta es dinámicamente establecido sobre el tiempo basado en la actividad y permanencia de cada uno. Este tiene un rol significativo durante la operación de consolidación (es decir, snapshot). Dado que los nodos corresponden con configuraciones limitadas de hardware, ellos no cuentan con suficientes recursos para mantener la historia completa de la cadena de bloques. Por esa razón, cada cierto volumen de operaciones, el nodo vigente con mayor puntuación puede solicitar la operación de consolidación la cual sintetiza toda la historia de la cadena de bloques en un único nodo con la información completa integrada para todos los nodos (es decir, se obtiene el último estado del registro derivado de toda su historia en una operación). Pareciera ser similar a un reinicio de la cadena de bloques, sin embargo, esta operación contiene la información del último nodo sin ninguna modificación. Esto es importante de mencionar en este contexto porque ayuda a liberar recursos en los nodos dada sus limitaciones. Es importante mencionar que la operación de consolidación solo la inicia el nodo con mayor puntaje, sin embargo, requiere de la aprobación de todos los nodos previo a avanzar en la consolidación o síntesis. En otras palabras, si los nodos no aprueban la operación, la operación de consolidación se deniega y la cadena de bloques se mantiene.

Así, cada transacción solicitada necesita ser aprobada por el resto de los nodos (Ver la clase *EndorsementDetails* en la Figura 57) y reunir consenso. Cuando el número de

nodos conectados es inferior a 11, se requiere el consenso del 100% de los participantes. Sin embargo, cuando el número de nodos excede los 10, se requiere la aprobación de todos los nodos que integran el 75% del mayor puntaje de acuerdo con la Ecuación 33.

La clase *LocalRecord* en la Figura 57 Principales Conceptos Relacionados con el Registro de Nodos Distribuidos basados en Blockchaincontempla las operaciones básicas descritas por los componentes *LocalResources, SenderLocalResources, y ReceiverLocalResources* (Ver Figura 56). Así, se mantiene seguimiento de las aprobaciones solicitadas y las decisiones alcanzadas por cada nodo. También se mantiene registro de las transacciones iniciadas hasta el instante en que se aprueban o no (Ver *Transactions* and *Transaction*). Como se mencionó anteriormente, la transacción puede contemplar una de cuatro operaciones:

- **Add**: Solicita la incorporación de un ítem de registro (o bloque).

- **Remove**: Solicita el borrado de su propio registro del registro distribuido (por ejemplo, podría darse cuando el nodo abandona la red de sensores).

- **UpdateRecordItem**: Se ejecuta cuando algún dato del ítem (o nodo) ha sido actualizado (por ejemplo, su ubicación).

- **Snapshot**: La comienza el nodo con mayor puntuación. Solicita la operación de consolidación sobre la cadena de bloques completa.

Cada adaptador o pasarela almacena un registro del ítem (Ver clase *Item*) en la base de datos distribuida, detallando su información. El ítem contiene un identificador de nodo (MA_ID), una lista separada por comas con las tarjetas de red autorizadas a transmitir medidas (*ANC*), una lista separada por comas de las fuentes de datos autorizadas (o sensores) para informar medidas (*ADS*), un mensaje basado en el lenguaje de marcado geográfico describiendo su posición actual (GML), y su clave pública. Esta información es útil especialmente porque permite conocer la proximidad entre nodos, pero también la similitud en términos de los sensores empleados, caracterizando las regiones monitoreadas.

Cada transacción de un nodo es almacenada en un árbol de Merkle para verificación de integridad (Ver *BDTree*). El árbol contiene la huella de cada transacción hasta que la operación de consolidación se ejecuta. Además, la huella de la raíz es indicativa de la integridad de las transacciones pasadas para un nodo y será incorporada como dato de la transacción cuando el nodo desee iniciar una nueva. Así, el árbol se actualiza cada vez que una nueva transacción se envía.

La cadena de bloques, el árbol de Merkle, y los registros de transacciones pueden ser regenerados completamente desde el registro distribuido. Por tal razón, incluso cuando

un nodo pierde su conectividad (o se queda sin batería), los registros podrían recrearse completamente solicitando un nuevo acceso al registro.

Dado que cada adaptador o pasarela posee las definiciones de proyectos en las que participa (ejemplo, mediante BriefPD) describiendo los nodos que la componen, cada nodo puede localizar nodos con los que se complementan sus proyectos de medición basado en la similitud estructural de la distancia compuesta ejecutada localmente a cada nodo. Además, puede capitalizar recomendaciones en las pasarelas asociadas con esos proyectos si requiriese tomar una decisión local (ejemplo, activar una sirena cuando la concentración de material particulado excede un umbral).

6.4.2 Perspectiva Comportamental

Dado que el registro unificado y distribuido de nodos basado en Blockchain se entiende para aplicarse en pasarelas y adaptadores, en sistemas de recolección de datos en tiempo real con hardware de capacidades limitadas, el comportamiento debiera respetar el principio de parsimonia del modelo [155]. De este modo, la funcionalidad debiera asegurar la confiabilidad e integridad del contenido de la base de datos, a sabiendas de las capacidades limitadas de los recursos [156].

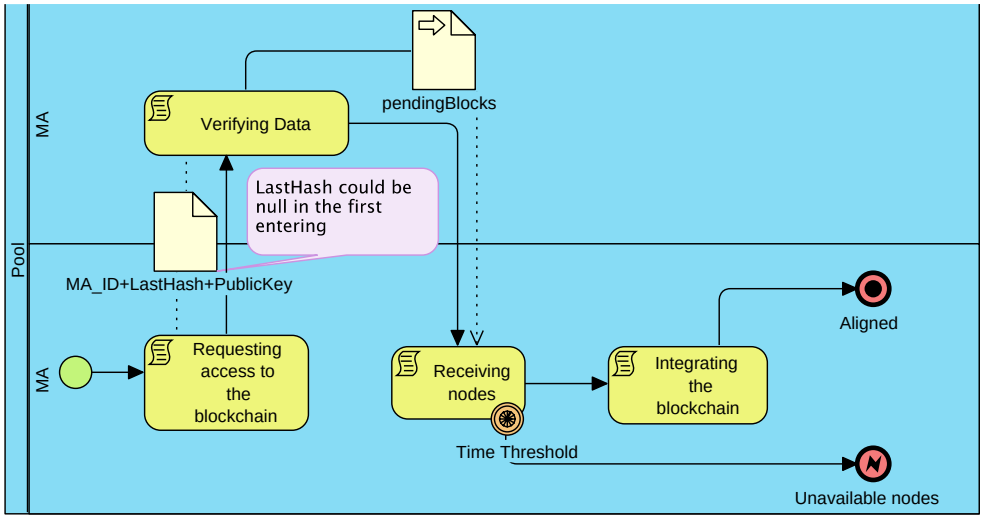

Figura 58 Diagrama BPMN describiendo la actualización de la base de datos distribuida

Tanto los adaptadores de medición como las pasarelas acceden a la base de datos distribuida, mientras que solo los adaptadores puede actuar sobre sus propios datos. La Figura 58 Diagrama BPMN describiendo la actualización de la base de datos distribuidadescribe los pasos relacionados con la actualización. El nodo que desea actualizar la base de datos inicia la operación adjuntando su ID, la huella del último nodo (cuando está disponible), y su clave pública. El adaptador de medición que recibe la solicitud, analiza la huella pública y la última huella del bloque de datos para determinar si el solicitante es un nodo válido o no (o es un componente foráneo). Esto es útil para determinar la prioridad de atención, primando los nodos registrados y colocando el

resto de las solicitudes al final de la cola. De acuerdo a la huella del nodo informado, se proveen los bloques de datos pendientes. Cuando el solicitante recibe el conjunto de datos pendiente, se los integra y consolida dentro de la cadena de bloques. Sin embargo, cuando el solicitante no recibe una respuesta luego de un tiempo dado, la solicitud es cancelada entendiendo que el nodo no está disponible. Esta situación podría ocurrir cuando no hay respuesta o cuando el nodo responsable de responder tiene otras solicitudes con mayor prioridad.

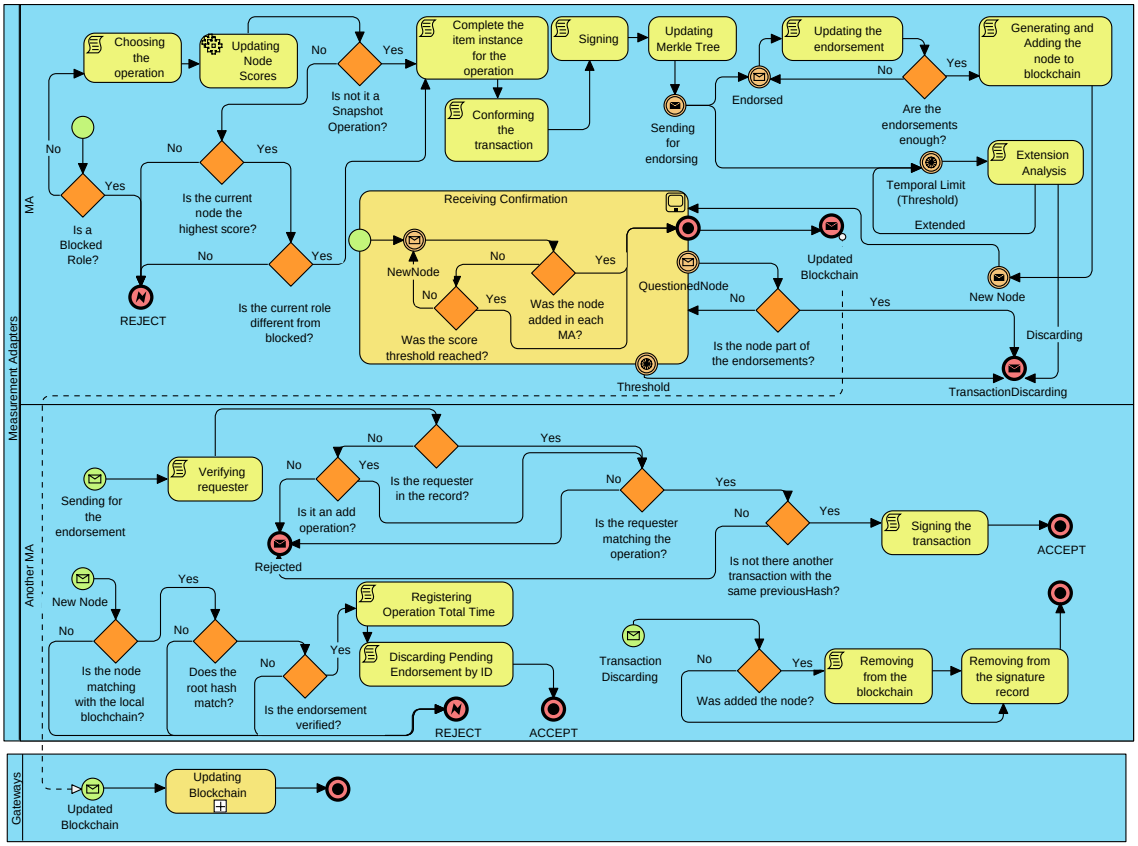

Figura 59 Un diagrama BPMN describiendo las Operaciones Principales relacionadas a la Base de Datos Distribuida

Sin embargo, no todas las operaciones pueden iniciarse desde cualquier nodo. La operación de consolidación (snapshot) es la excepción. Esta es iniciada por el nodo de mayor puntuación en base a la fórmula Ecuación 33 (Ver área superior izquierda de la Figura 59 Un diagrama BPMN describiendo las Operaciones Principales relacionadas a la Base de Datos Distribuida),el acercamiento de Bitcoin en relación al concepto de prueba de trabajo [157]. Dado que el cómputo de la puntuación es dinámica, cada vez que un nodo solicita una transacción se actualizan las puntuaciones de los nodos.

El resto de las operaciones pueden ser iniciadas por cualquier nodo registrado, tomando una serie de acciones limitadas a sus propios datos exclusivamente. Cuando un adaptador o pasarela no se encuentra registrado, puede iniciar una operación *add* para incorporar sus datos en la base de datos. Luego de ello, el nodo es responsable por mantener actualizados sus datos.

Cuando se inicia una transacción con una o más operaciones, ésta debe ser firmada por el solicitante utilizando su clave privada. Luego de ello, se envía la transacción a los restantes nodos de la base de datos para su aprobación. Cada nodo inicia su proceso de aprobación al recibir el mensaje. Allí, se verifica al solicitante para saber si está presente en la base de datos. Si la operación solicitada es diferente de *add* y el solicitante no está presente, no se aprueba la operación. De otro modo, si el solicitante está presente o la transacción se asocia con una operación *add*, se verifica la transacción para verificar la integridad entre transacción y solicitante. Cuando otra transacción previa tiene la misma huella para el bloque de datos previo, la transacción actual no se aprueba (esto implica que se generó a destiempo). Sin embargo, cuando la transacción empareja con la huella del último bloque de la cadena, la transacción se aprueba.

El solicitante espera por el arribo de las aprobaciones. Cuando la suma de las puntuaciones de las aprobaciones recibidas es superior o igual al umbral de autorización, la transacción recibe la aprobación final. Luego de ello, el solicitante crea el nuevo nodo con los datos aprobados, lo distribuye e integra en la cadena de bloques. Sin embargo, cuando no se logra la autorización el solicitante espera por un tiempo dado. Si el tiempo de espera se excede, la transacción se marca para descarte. Cuando un nodo recibe una solicitud de descarte, remueve la transacción de la cadena de bloques y del registro local.

Por otro lado, cuando se genera un nuevo nodo, este se comunica mediante el mensaje *AddedNode*. En este caso, el solicitante actúa como coordinador, esperando por los mensajes de respuesta dentro de los restantes nodos. Cuando el solicitante alcanza la puntuación necesaria, la transacción se torna en aprobada independientemente si aún restan respuestas. Esto es para evitar situaciones en las cuales nodos registrados no provean respuestas porque están fuera de servicio, por ejemplo, por fallas de energía. Incluso más, si una transacción alcanza el umbral de aprobación, no importa que otro nodo cuestionó al nuevo nodo, será ratificado de todos modos. El árbol de Merkle contiene cada transacción enviada desde un nodo para responder las consultas de integridad desde el resto de los nodos.

Sin embargo, cuando el número de puntuaciones no permite lograr un umbral en un tiempo dado, la transacción se descarta. De este modo, no será incorporada en la cadena de bloques.

6.4.3 Implementación de Referencia

A los efectos de materializar los conceptos y propuestas esgrimidas, se ha desarrollado una implementación de referencia utilizando Java con microservicios a través del marco Spring Boot 2.5.0. La librería con el código implementado está disponible en la librería nrchain en GitHub bajo los términos de la licencia 2.0 [156].

Se ejecutaron dos simulaciones discretas sobre una MacBook Pro con 16GB de RAM, MacOS Big Sur 11.4 y un procesador Intel Core i7 de 2.9 Ghz. Ambas simulaciones

emplean com.github.jbellis.jamm package (version 0.3.3) como agente instrumental para medir el consumo de memoria. El objetivo de las simulaciones tuvo las siguientes finalidades:

- **Simulación 1**: Analizó el tamaño de memoria requerido cuando el número de nodos crece de 1 a 200 sin operación de consolidación (snapshot). El primer nodo se utilizó como referencia para la inicialización de la cadena de bloques respecto de los sucesivos nodos. Una vez creado el nodo, la cadena de bloques es actualizada a través del nodo de referencia, luego de lo cual, el nodo se auto agrega en la cadena de bloques. La operación es repetida hasta alcanzar los 200 nodos (Este representa un parámetro arbitrario y puede ser modificado). Cuando el número máximo de nodos es alcanzado, cada nodo se apaga y las medidas se almacenan en un archivo consolidado.

- **Simulación 2**: Esta simulación es similar a la anterior, aunque establece un parámetro "each" que determina la realización de una operación de consolidación (snapshot) cada cierto número de nodos incorporados. En este caso, ese valor se define en 20.

De este modo, la idea de las simulaciones es contrastar el efecto en el consumo de memoria cuando se incorpora la operación de consolidación de cuando se prescinde de ella.

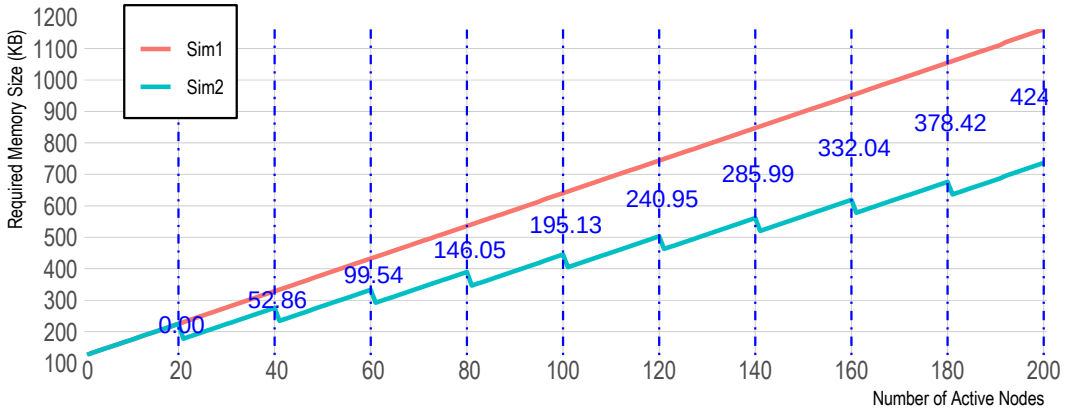

Figura 60 Evolución del Consumo de Memoria, incorporando consolidación (simulación 2) y prescindiendo (simulación 1) de ella.

Como puede apreciarse en la Figura 60 Evolución del Consumo de Memoria, incorporando consolidación (simulación 2) y prescindiendo (simulación 1) de ella., a partir del instante en que comienza la operación de consolidación (cada 20 nodos) se observa una diferenciación en el consumo de memoria. Solo entre 20 y 39 nodos la diferencia de consumo es alrededor de 52 kB. Entre 40 y 59 asciende la diferencia alrededor de 99 kB. Luego de ello, las diferencias van progresivamente acrecentándose con cada nueva operación a 146 kB, 195 kB, 240 kB, hasta alcanzar 424 kB con 200 nodos. Ello representa un ahorro importante ya que va de la mano con el incremento

de la cantidad de nodos. Por ejemplo, la simulación 1 indicó 1.161 kB por nodo para un registro de 200 nodos versus 737 kB en la simulación 2, lo que representa un ahorro del 36,52% en memoria.

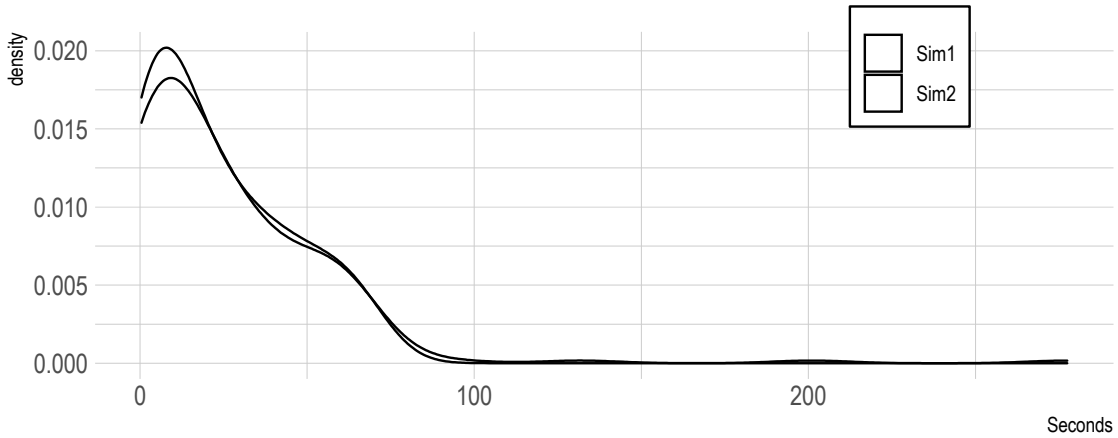

Figura 61 Curvas de Densidad para los Tiempos de Inicialización de Registro de la Simulación 1 y 2

La Figura 61 Curvas de Densidad para los Tiempos de Inicialización de Registro de la Simulación 1 y 2 describe las curvas de densidad para los tiempos involucrados en la inicialización de los nodos para ambas simulaciones. Como se puede apreciar, la simulación 1 tiene una sutil concentración a valores más bajos respecto de la simulación 2. Ello es esperado debido a la sobrecarga producida por la operación de consolidación en el tiempo de inicialización.

Para analizar la magnitud de las diferencias, la distancia de Hellsinger permite analizar la distribución cuantificando sus diferencias entre distribuciones de probabilidad. Valores cercanos a cero indicarían poca diferencia, mientras que aquellos cercanos a uno indicarían notables diferencias. En este caso, utilizando el paquete statip (versión 0.2.3) con los datos de simulación en R (3.6.2) se obtuvo 0,1047 lo cual es consistente con las curvas de densidad expuestas.

La Figura 62 Diagramas de Violín del tiempo de inicialización para los nodos en las simulaciones 1 y 2 describe las gráficas de violín para las simulaciones 1 y 2 respecto de los tiempos de inicialización de los nodos. A partir de ellos, es posible apreciar la concentración de sus valores junto con su distribución. Puede visualizarse en estos gráficos una distribución de datos similares, con pequeñas variaciones en la simulación 2 que producen un desplazamiento de la media aritmética.

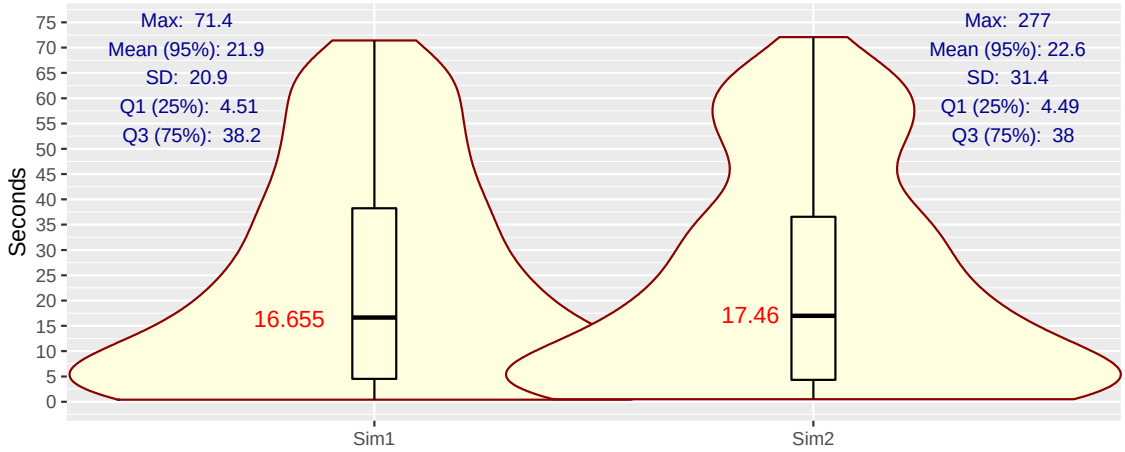

Figura 62 Diagramas de Violín del tiempo de inicialización para los nodos en las simulaciones 1 y 2

Independientemente de las diferencias mencionadas, los rangos inter cuartiles y su constitución son bastante similares visualmente. Sin embargo, debido a valores atípicos derivados por el tiempo adicional requerido por la operación de consolidación (simulación 2), la media aritmética se ve afectada. Así, el peor escenario para la simulación 2 fue que un nodo requirió 277 segundos para inicializarse, mientras que en la simulación 1 fue de 71,4 segundos. Sin embargo, es importante mencionar que la sincronización de la cadena de bloques desde el nodo de referencia se incorpora al tiempo de inicialización y de allí la diferencia de magnitudes.

6.5 Conclusiones Generales del Capítulo

El capítulo cuatro tecnologías complementarias para mejorar la PAbMM y fortalecer el cálculo de la distancia compuesta como conductor de la búsqueda de recomendaciones. Se introdujo la posibilidad de detectores de cambio y barreras temporales para focalizar en los datos de mayor impacto en el análisis comportamental entre proyectos. Se abordó el descarte selectivo como herramienta de priorización de las métricas esenciales a la entidad bajo monitoreo y su contexto. Se abordó el uso de árbol de Merkle para mejorar la verificación de integridad de las medidas transmitidas incrementado. Finalmente, se abordó el uso de cadena de bloques para descomprimir la dependencia de la nube, fortalecer acceso a datos y recomendaciones próximas al nodo, de modo que pueda tomar decisiones localmente.

Los detectores de cambios y barreras temporales junto con la organización del búfer de datos permiten focalizar sobre un subconjunto de datos con suficiente prueba estadística para producir un cambio. Esto tiene impacto directo sobre el cálculo de la similitud comportamental e incorpora un mecanismo de revisión en el origen del dato de particular importancia. Para monitorear 100 métricas simultáneamente se consumiría 542,05 kB antes de procesar datos y 568,09 kB con las estimaciones incorporadas (luego de procesar datos). Así, los datos relacionados con las estimaciones consumirían 26,04 kB en 100 métricas. La mediana del tiempo para la operación de

determinar la presencia (o no) de cambios consumió 238 ns (0,000238 ms), mientras que la mediana para la operación de agregación de medidas requirió 918 ns (0,000918 ms). La organización del búfer implementado una ventana lógica permite mantener las últimas medidas coordinadamente con los detectores de cambio y minimizando riesgos de desborde de memoria ya que es posible establecer a priori un umbral.

Las técnicas de descarte selectivo junto con los detectores de cambio permitirían evitar hasta un 74,57% de las transmisiones de datos, consumiendo solo los recursos asociados con un 25,43% de las transmisiones originales.

El registro de integridad basado en árbol de Merkle permite contrastar en una sola operación la integridad de las transmisiones de medidas realizadas sin requerir almacenar su contenido. La operación más costosa desde el punto de vista temporal fue *addTransaction* con una mediana de 0,0666 ms y una media de 0,0767 ms. Ello incorpora el tiempo total de recálculo de huellas a lo largo de la jerarquía.

El registro distribuido es una herramienta importante para descentralizar la información de los nodos, intercambiar definiciones de proyectos localmente (sin requerir de la nube) y poder calcular similitudes entre proyectos junto con búsquedas de recomendaciones localmente a partir de la distancia compuesta. El pero tiempo de inicialización con operaciones de consolidación implicó 277 segundos, aunque el ahorro en el tamaño requerido de memoria para los nodos implicó un ahorro del 36.52%.

Estas tecnologías de soporte de PAbMM permite mejorar la cualidad de los resultados de la distancia compuesta, incluso en instancias de stress de procesamiento por cuanto focalizan en la confiabilidad de las medidas, retener aquellas estadísticamente plausibles de producir cambios, y descentralizando los registros para fomentar el cómputo local de distancia compuesta por nodo. Este capítulo abordó tecnologías complementarias con impacto positivo en el cálculo de la distancia responsable de la priorización de proyectos por similitud.

Capítulo 7	Escenario de Uso. Monitoreo de Material Particulado

Capítulo 7. Escenario de Uso. Monitoreo de Material Particulado

Introducción

El capítulo anterior introdujo diversas tecnologías de soporte a la arquitectura de procesamiento basada en metadatos de mediciones con impacto en la distancia compuesta. Entre las cuatro principales mejoras incorporadas a PAbMM, se mencionaron:

- Los detectores de cambio y barreras temporales. Estos permiten optimizar la política de transmisión de datos, pero a su vez, asegurar que los datos informados a partir de los detectores son suficientes estadísticamente para justificar un cambio en el dato monitoreado. Este es un aspecto esencial desde la perspectiva comportamental de la distancia compuesta.

- El descarte selectivo basado en puntuaciones Z. Esto permitió incorporar la implementación de una ventana lógica a nivel de búfer para evitar desbordes en los adaptadores a la vez que siempre se informan las últimas medidas conocidas. A su vez, en caso de ser necesario priorizar medidas, se analizó la priorización de métricas como elementos de referencia para retener datos. Este aspecto es importante para asegurar que ante situaciones límites (por ejemplo, exceso de la capacidad de procesamiento), es posible guiar a los adaptadores sobre la importancia de las métricas y medidas a transmitir. De la simulación discreta pudo observarse que los detectores de cambio se complementaban muy bien con el balance de carga evitando el descarte, lo que sumado a la prioridad de las métricas permite el aprovisionamiento de datos de interés a la distancia compuesta.

- El registro de integridad mediante árbol de Merkle. Este avance permitió incorporar un registro de integridad a nivel de PAbMM pero también a nivel del adaptador de medición. Ello le permite tener un seguimiento y control de integridad de las transacciones recibidas de los adaptares, pero también, permite a PAbMM verificar contra el propio adaptador empleándolo como segundo factor.

- Registro Unificado de Nodos basado en Blockchain. El registro unificado de nodos es empleado tanto a nivel de PAbMM como de nodos para conocer la ubicación de los elementos recolectores y su proximidad. Esto es clave para determinar una transmisión indirecta a través de un tercer nodo. Sin embargo, el registro se encontraba centralizado en PAbMM y ante una pérdida de conectividad, el nodo quedaba aislado. Este mecanismo, permite que los nodos mantengan una base de datos distribuida sobre sus datos en el campo de

monitoreo. En caso de pérdida de conectividad con PAbMM, pueden conocer sus nodos cercanos e intentar pedir soporte a ellos para realizar las respectivas transmisiones.

Este capítulo introduce un escenario de uso para la arquitectura de procesamiento basado en metadatos de mediciones focalizado en el material particulado.

El capítulo se organiza en cinco secciones. La primera sección se centra en introducir el concepto de material particulado, su impacto potencial en la salud de las personas, y la importancia de su monitoreo. La segunda sección sintetiza los acercamientos, dispositivos, y estrategias actuales para monitorear el material particulado. La tercera sección introduce la descripción del escenario de uso en la provincia de La Pampa, mientras que la cuarta sección sintetiza la aplicación de la distancia compuesta. Finalmente se proveen conclusiones al respecto.

El capítulo se soporta en las siguientes publicaciones efectuadas a lo largo del proceso de investigación:

- Diván, M. Sánchez-Reynoso, M. Gonnet, S. (2022) **"Measurement Project Interoperability for Real-time Data Gathering Systems**". Future Generation Computer Systems, Elsevier, ISSN 0167-739X, 129, 298-314 https://doi.org/10.1016/j.future.2021.11.031

- Diván, M. Sánchez-Reynoso, (2021). **"Big Data Analysis for Green Computing: Concepts and Applications"**. Effect of the measurement on Big Data Analytics. An evolutive perspective with Business Intelligence. ISBN 9781003032328. pp 50-69. Estados Unidos. CRC Press. http://dx.doi.org/10.1201/9781003032328-4

- Diván, M. Sánchez Reynoso, M. Méndez, M. Panebianco J. (2021) "**IoT-based Approaches for Monitoring the Particulate Matter and its Impact on Health"**. e-ISSN: 2327-4662 – IEEE Internet of Things Journal. Institute of Electrical and Electronics Engineers Inc. 2021. Vol. 8, nro 15. pp. 11983 - 12003 http://dx.doi.org/10.1109/JIOT.2021.3068898

- Diván, M. Sánchez Reynoso, M. (2021) **"Strategies based on IoT for supporting the decision making in Agriculture: A Systematic Literature Mapping**". ISSN: 1755-0556 - e-ISSN: 1755-0564 - International Journal of Reasoning-based Intelligent Systems. Vol. 13, nro 3. pp155 – 171. https://dx.doi.org/10.1504/IJRIS.2021.117080

- Diván, M. Sánchez-Reynoso, M. Gonnet, S. (2021) **"Recent Applications of Federated Learning in Edge and IoT Environments: A Review**". Proceeding of

the 2021 5th International Conference on Information Systems and Computer Networks (ISCON). IEEE. https://doi.org/10.1109/ISCON52037.2021.9702466

7.1 El Material Particulado y su Impacto Potencial en la Salud

El material particulado en el aire se constituye por la totalidad de las partículas (sólidas y líquidas) suspendidas en el aire. Su composición podría variar dependiendo de la región, mientras que poseen un tamaño estimativo entre unos pocos nanómetros y 100 micrómetros [158]. Se lo refiere comúnmente por su acrónimo en inglés PM (Particulate Matter).

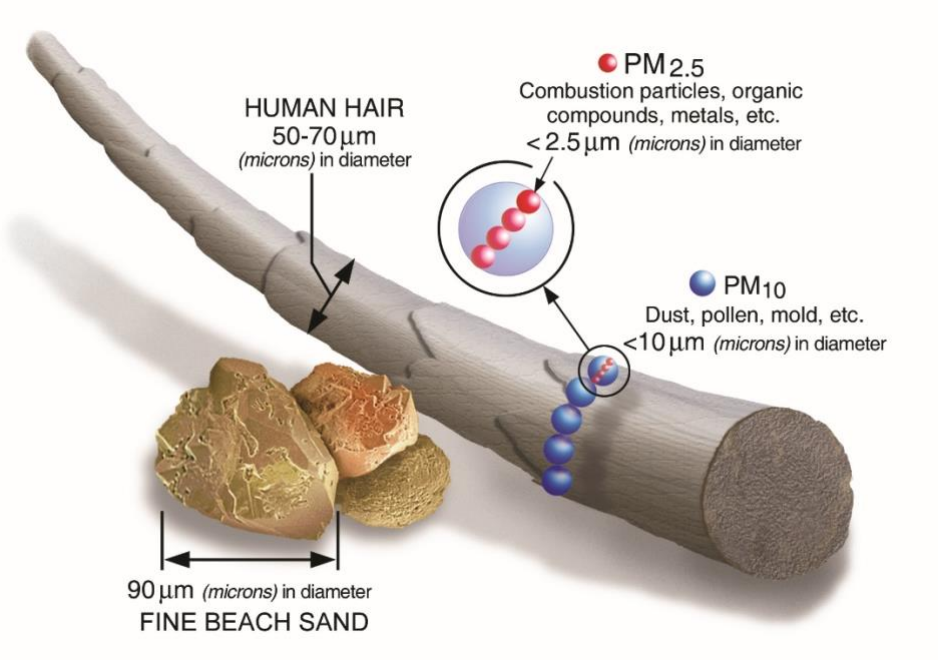

Figura 63 Comparación de Tamaños para el Material Particulado. Fuente: Environmental Protection Agency, United States of America

La Figura 63 describe una perspectiva comparativa del tamaño del material particulado. En ella puede apreciarse 1) La arena fina de playa con diámetros cercanos a los 90 micrones (en inglés, *Fine Beach Sand*); 2) El pelo humano (en inglés, *Human Hair*) con diámetros entre 50 y 70 micrones; 3) El polvo (en inglés, *Dust*), polen (en inglés, *Pollen*), moho (en inglés, *Mold*), entre otros con diámetros inferiores a los 10 micrones. Partículas de este tipo se refieren comúnmente como PM10 en alusión a que son todas aquellas cuyo diámetro se encuentra debajo de los 10 micrones; 4) Partículas de combustión (en inglés, *Combustion Particles*), compuestos orgánicos (en inglés, *Organic Compounds*), metales, entre otros con diámetros inferiores a 2,5 micrones. Partículas de este tipo se refieren comúnmente como PM2.5 en alusión a que son todas aquellas cuyo diámetro se encuentra debajo de los 2.5 micrones.

De acuerdo con las proyecciones de mortalidad y causas de muerte de la Organización Mundial de la Salud (OMS) actualizado a 2016, las enfermedades respiratorias ocupan la cuarta posición en el ranking mundial de causas de muerte, mientras que la polución del aire puede asociarse con alrededor del 8% de las muertes anuales [159], [160].

Debido al tamaño asociado con el material particulado, este podría invadir diferentes órganos del cuerpo produciendo diferentes enfermedades, afectando el tracto respiratorio, o influenciando respecto de enfermedades prexistentes [161]–[163]. Por esta razón, se han desarrollado diversas estrategias para estudiar y monitorear la calidad del aire, analizando el material particulado en concentración y composición [164]–[166].

El Internet de las Cosas emergió como una alternativa para implementar un gran número de sistemas de recolección de datos en tiempo real [167]. Es decir, la posibilidad de contar con dispositivos accesibles, pequeños, baratos y disponibles permitió implementar diferentes aplicaciones en un amplio rango de países [168], [169]. En este tipo de contexto, los datos necesitan ser procesados tan pronto como arriban para soportar un proceso de toma de decisiones en tiempo real [170]. La capacidad de reaccionar instantáneamente ante el arribo del dato constituye un activo clave para el monitoreo de la calidad del aire, la detección de material particulado y su composición [171].

Las características y composición del material particulado son afectadas por el contexto, las actividades desarrolladas en la región junto con el clima (entre otros factores). De este modo, las estrategias de medición basadas en Internet de las Cosas podrían abordarse de diferentes modos de acuerdo con el objetivo perseguido [158], [163], [164]. Dada la variedad de acercamientos, tanto en término de hardware como de diseño experimental, la siguiente sección sintetiza las características generales de las estrategias de medición implementadas. Especialmente, aquellas estrategias que emplean sensores y placas de procesamiento basadas en Internet de las Cosas para estudiar los efectos del material particulado en la salud de las personas.

7.2 Estrategias actuales de monitoreo

Esta sección sintetiza nuestro estudio sistemático de la literatura conducido sobre las bibliotecas digitales de ACM, IEEE, Scopus, Springer Link, Science Direct, y Wiley publicado en 2021 para conocer el escenario y aproximaciones basadas en Internet de las Cosas orientadas al monitoreo de material particulado en aplicaciones relacionadas con la salud de las personas [172]. De este modo, la primera parte de la sección introduce sintéticamente la metodología empleada para llevar adelante el estudio. La segunda parte describe principales hallazgos respecto de las dimensiones y características definidas. Finalmente, el tercer apartado sintetiza los principales resultados y conclusiones del estudio.

7.2.1 Metodología

Esta sección describe la metodología aplicada para conducir un estudio sistemático de la literatura con el objetivo de proveer una perspectiva amplia y consolidada del uso de los sistemas basados en Internet de las Cosas para el monitoreo de materia particulado con aplicación en la salud de las personas y esquemas de recomendación asociados.

Las preguntas de investigación derivadas a partir del objetivo indicado son sintetizadas en la Tabla 31 en el idioma original en que se plantearon en el estudio y su correspondiente traducción al español. Es necesario mantener las preguntas de investigación en inglés, dado que las dimensiones y palabras claves derivadas para formular las consultas en las bases de datos se obtienen a partir de ellas y se ejecutan en inglés.

Tabla 31 Preguntas de Investigación

RQ	Pregunta Original	Pregunta en español
1	What are IoT-based measurement approaches implemented to monitor PM?	¿Cuáles son los enfoques basados en Internet de las Cosas implementadas para monitorear material particulado?
2	How are the consistency and repeatability of the IoT measurement process warrantied over time?	¿Cómo se garantiza en el tiempo la consistencia y repetibilidad del proceso de medición basado en Internet de las Cosas?
3	How are the collected measures used to evaluate the impact of PM on human health?	¿Cómo se usan las medidas obtenidas para evaluar el impacto del material particulado en la salud de las personas?
4	How is the decision-making process articulated with a real-time collecting system to monitor PM?	¿Cómo se articula el proceso de toma de decisiones con un sistema de recolección en tiempo real para monitorear material particulado?
5	What is kind of IoT-based devices used to monitor PM in each application domain?	Cuál es el tipo de dispositivo utilizado para monitorear material particulado en cada dominio de aplicación.

La pregunta de investigación (en inglés, Research Question -RQ) 1 tiene por finalidad explorar diferentes alternativas y estrategias para estudiar material particulado utilizando dispositivos de Internet de las Cosas, independientemente del dominio de aplicación.

La segunda pregunta focaliza en analizar la asociatividad entre las medidas obtenidas, los dispositivos, y el proceso de medición. Es decir, independientemente del enfoque empleado para articular los dispositivos con los esquemas de recolección de datos, el punto refiere a cómo se asegura la comparabilidad de las medidas en el tiempo.

La tercera pregunta se orienta a conocer de qué modo las medidas se emplean para evaluar (directa o indirectamente) la incidencia en la salud de las personas. Es decir, la

idea es conocer de qué modo se emplean para estudiar, discriminar, identificar, o anticipar cualquier tipo de efectos nocivos en la salud.

La cuarta pregunta se centra en el proceso de toma de decisiones. En otras palabras, la inquietud se centra en conocer el acercamiento empleado para articular los sistemas de recolección en tiempo real con respecto a la capacidad de prevenir, actuar, o tomar una decisión basado en las medidas.

La quinta pregunta de investigación se orienta a conocer las características de los enfoques empleados por cada alternativa de monitoreo de material particulado de acuerdo con el escenario o entorno en donde se aplica.

Tabla 32 Palabras Claves derivadas de las Preguntas de Investigación

RQ	Palabras Claves
1	IoT (alternatively “Internet-of-Things” or “Internet of Things”), Measurement, Approach, Monitor, Particulate Matter (or PM)
2	Consistency, Repeatability, IoT, Measurement, Process
3	Measures, Evaluate, Impact, Particulate Matter, Health
4	Decision-making, Process, Real-time, Collecting, Monitor, Particulate Matter
5	Kinds (alternatively “Types”), IoT, Device, Monitor, Particulate Matter

La Tabla 32 describe el conjunto de palabras claves derivadas a partir de las preguntas de investigación. A partir de la combinación de estas se formula la cadena de búsqueda a ejecutar en los repositorios digitales. Sin embargo, dicha consulta debe ser ajustada de acuerdo con cada repositorio dado que cada uno posee una sintaxis particular. La Tabla 33 presenta las cadenas de búsqueda ajustadas por repositorio digital.

Tabla 33 Cadenas de Búsqueda por Repositorio Digital

Repo	Cadena de Búsqueda
ACM	[[All: "internet-of-things"] OR [All: "internet of things"] OR [All: "iot"]] AND [[All: "particulate matter"] OR [All: "pm"]] AND [All: "health"] AND [[All: "measurement"] OR [All: "approach"] OR [All: "monitor"] OR [All: "consistency"] OR [All: "repeatability"] OR [All: "process"] OR [All: "measures"] OR [All: "evaluate"] OR [All: "impact"] OR [All: "decision-making"] OR [All: "process"] OR [All: "real-time"] OR [All: "collecting"] OR [All: "kinds"] OR [All: "types"] OR [All: "device"]]
IEEE	("All Metadata":"Internet-of-Things" OR "All Metadata":"Internet of Things" OR "All Metadata":"IoT") AND ("All Metadata":"Particulate Matter" OR "All Metadata":"PM") AND "All Metadata":"Health" AND ("All Metadata":"Measurement" OR "All Metadata":"Monitor" OR "All Metadata":"Consistency" OR "All Metadata":"Repeatability" OR "All Metadata":"Process" OR "All Metadata":"Measures" OR "All Metadata":"Evaluate" OR "All Metadata":"Impact" OR "All Metadata":"Decision-making" OR "All Metadata":"Real-time" OR "All Metadata":"Collecting" OR "All Metadata":"Kinds" OR "All Metadata":"Types" OR "All Metadata":"Device")

Springer Link	Internet AND of AND Things AND Particulate AND Matter AND Health AND (Measurement OR Approach OR Monitor OR Consistency OR Repeatability OR Measures OR Evaluate OR Impact OR Decision-making OR Process OR Real-Time OR Collecting OR Kinds OR Types OR Device)
Science Direct	Artículos con los siguientes términos: ("Measurement" OR "Monitor" OR "Repeatability" OR "Process" OR "Evaluate" OR "Impact" OR "Decision-making" OR "Real-time" OR "Device") Título, resumen, palabras claves: ("Internet-of-things" OR "Internet of things" OR "IoT") AND ("Particulate Matter" OR "PM") AND "Health"
Scopus	TITLE-ABS-KEY(("Internet-of-things" OR "Internet of things" OR "IoT") AND("Particulate Matter" OR "PM") AND "Health" AND("Measurement" OR "Approach" OR "Monitor" OR "Consistency" OR "Repeatability" OR "Process" OR "Measures" OR "Evaluate" OR "Impact" OR "Decision-making" OR "Real-time" OR "Collecting" OR ("Kinds" OR "Types") OR "Device"))
Wiley	"("Internet-of-things" OR "Internet of things" OR "IoT")" anywhere and "("Particulate Matter" OR "PM")" anywhere and ""Health"" anywhere and "("Measurement" OR "Approach" OR "Monitor" OR "Consistency" OR "Repeatability" OR "Process" OR "Measures" OR "Evaluate" OR "Impact" OR "Decision-making" OR "Process" OR "Real-time" OR "Collecting" OR ("Kinds" OR "Types") OR "Device")" anywhere

De este modo, se ejecutaron las consultas en los respectivos repositorios en septiembre de 2020. Se filtraron los resultados para limitar a artículos publicados en revistas y escritos en inglés. Se consolidó el registro de resultados en un archivo Excel y se eliminaron registros duplicados (dado que al provenir de distintas fuentes era un factor para considerar).

7.2.2 Dimensiones y Características

A partir de las preguntas de investigación, la Tabla 34 define las dimensiones y categorías de análisis. Estas son de especial utilidad para proveer una síntesis a partir de los resultados de las consultas.

Las consultas a los repositorios arrojaron 568 artículos. De ellos, se removieron resúmenes, índices, noticias, contenido no relacionado (solo mencionaban una palabra clave a modo de ejemplo), actas de conferencias, y revisiones. De la lectura individual de los trabajos resultantes de las cadenas de búsqueda y aplicados los respectivos filtros se retuvieron 55 artículos. De ellos, 7 se encontraban duplicados y se eliminaron, quedando 48. Estos se clasificaron en las categorías indicadas en la siguiente tabla para las dimensiones bajo análisis.

Tabla 34 Categorías y Dimensiones de Análisis

RQ	Dimensión	Categorías
1	Approaches	Focused; Distributed; Global
2	Measurement Process	Non-formalized; Formalized
3	Impact on Health	PM2.5; PM10; PM10+
4	Decision-making	Recommendation; Real-time; Offline; Use of Knowledge
5	Kind of devices	Laser-based; Infrared-based; Another method

En relación con la RQ1, los enfoques fueron 1) **Focalizados** (*focused*): cuando la estrategia se aplica a una región limitada (por ejemplo, un hogar); 2) **Distribuida** (*Distributed*) cuando la estación de monitoreo trabajó colaborativamente con otras en un límite dado (ejemplo, una ciudad); 3) **Global**: cuando las estaciones de monitoreo se articulan a lo largo del planeta. 33,33% de los trabajos retenidos correspondieron a enfoques focalizados, 64,70% a distribuidos, y 1,97% a globales. La motivación alrededor de los enfoques globales y distribuidos reside en incrementar la cobertura y densidad para el monitoreo ambiental basado en dispositivos de bajo coste dado la elevada inversión que implica un equipamiento de alta exactitud como los empleado por las agencias de protección ambiental de los gobiernos.

En relación con la RQ2, se discrimina entre 1) **Formalizado** (*Formalized*): cuando el proceso de medición se basa en algún marco formal de medición; 2) **No Formalizado** (*Non-formalized*): cuando el proceso de medición se implementa ad-hoc. El 64,44% de los artículos revisados poseen un proceso de medición formalizado, es decir, que de algún modo describen apropiadamente el proceso de calibración, el equipamiento empleado para calibrar, y algún modo de ajustar las medidas obtenidas desde los sensores de bajo costo. El 35,56% de los trabajos no mencionan proceso de calibración alguno, parecieran emplear los sensores en forma directa lo que sería un factor de sesgo para los valores.

La RQ3 focalizó en el tamaño de material particulado analizado por los diferentes trabajos, habida cuenta del impacto potencial en la salud a partir de este. 1) **PM2.5**: Refiere a estudios que abordan el material particulado hasta 2,5 micrones; 2) **PM10**: Refiere a estudios que abordan el material particulado (2,5; 10] micrones hasta 10 micrones; 3) **PM10+**: Se vincula con estudios que estudian material particulado mayor a 10 micrones. El 59,15% de los artículos analizados abordan PM2.5 y el el 40,85% refiere a PM10. Ninguno analizó tamaños de partículas por encima de los 10 micrones, lo cual es esperable dado el impacto en la salud de los primeros.

La RQ4 describe las categorías asociadas con el modo en que se usa o complementa las medidas para tomar las decisiones. 1) **Recomendación** (*Recommendation*): se relaciona con trabajos que proveen sugerencias o implementan acciones a partir de una decisión tomada en base a las medidas; 2) **Tiempo Real** (*Real-time*): Se refiere a los trabajos que toman decisiones tan pronto como el dato arriba; 3) **Fuera de línea**

(*Offline*): En este tipo de enfoques la toma de decisiones ocurre en segundo plano o dispone de suficiente tiempo para analizar los datos sin proveer una respuesta inmediata; 4) **Uso del conocimiento** (*Use of Knowledge*): En estos casos la toma de decisiones se soporta por algún tipo de conocimiento de expertos o basado en experiencias previas, independientemente de si se proveen recomendaciones o no. Aquí debe mencionarse que un artículo podría emplear más de una categoría en simultáneo, por ejemplo, tomar decisiones en tiempo real y proveer recomendaciones. Por ello, los 48 artículos fueron incorporados en tantas categorías como corresponde y los porcentajes refieren al total (48). Así, se tiene que el 12,50% de los artículos introduce algún esquema de recomendación, 85,41% direcciona el procesamiento de datos en tiempo real, 43,75% incorpora algún modo de análisis en segundo plano con procesamiento por lotes, y el 10,41% considera el empleo de experiencias previas o conocimientos.

La RQ5 se centró en el tipo de método empleado para detectar y medir el material particulado. 1) **Basado en láser** (*Laser-based*); 2) **Basado en Infrarrojo** (*Infrared-based*); 3) **Otro método** (*Another pertinente method*). Los dispositivos empleados en los artículos corresponden a un 38% basados en láser, 34% basado en infrarrojo, 6% emplea tecnología ad-hoc, y 22% no los especifica. La Figura 64 describe un diagrama de embudo donde ordena los principales sensores para detectar material particulado empleado en los trabajos analizados.

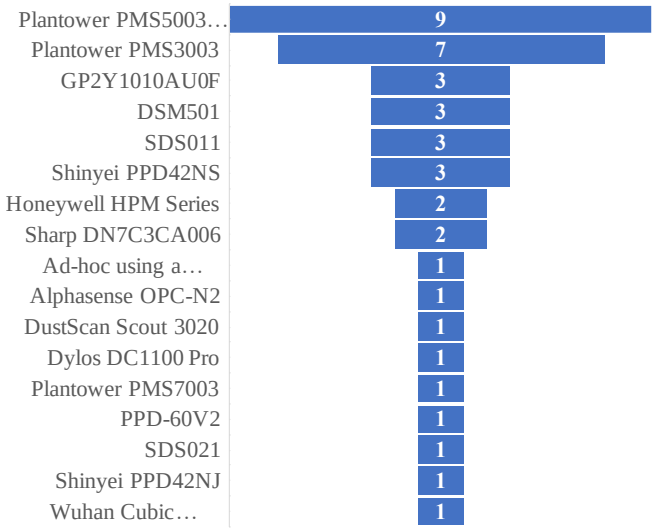

Figura 64 Principales Sensores para Detectar Material Particulado

Como puede observarse en la anterior figura, las opciones populares para detectar material particulado son el Plantower PMS 5003 y 3003. El primero es un sensor basado en láser que puede detectar partículas entre [1; 10] micrones e informar la masa de la partícula con un precio estimado de 35 USD. El segundo es un sensor basado en láser, con un rango similar de detección de material particulado, pero sin estimación de la concentración de la masa y un precio estimado de 25 USD.

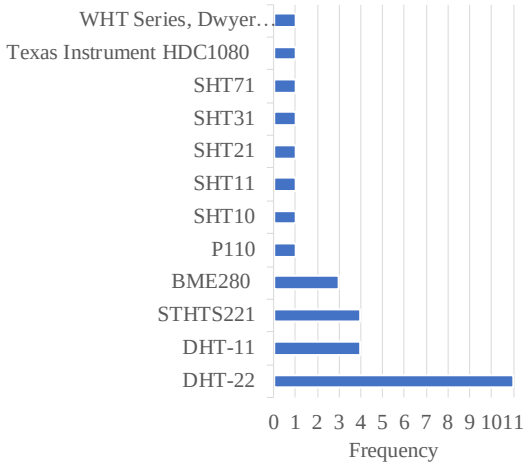

Figura 65 Sensores de Temperatura y Humedad Relativa

La Figura 65 describe los sensores de humedad relativa y temperatura utilizados en los artículos. Estos se incorporan junto con el sensor de material particulado dado que la variación de temperatura y humedad afectan sensiblemente las mediciones de material particulado [173]. El sensor más común fue el modelo DHT-22 con un precio estimado de 7 USD.

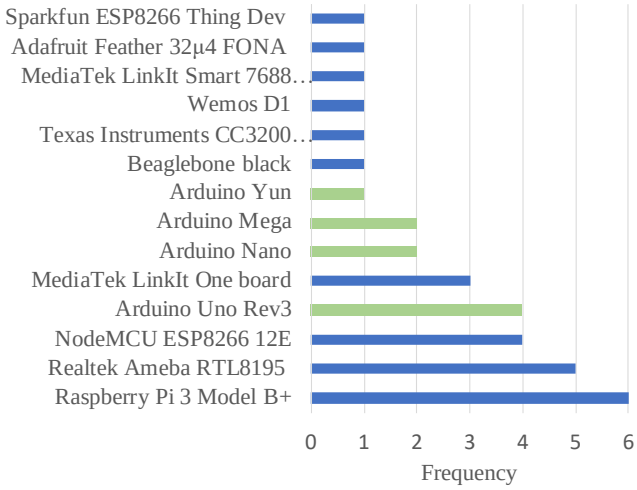

Figura 66 Principales Computadoras de Placa Simple Empleadas

La Figura 66 describe las principales computadoras de placa simple (*Single Board Computer -SBC*) empleados en diferentes proyectos de monitoreo. La opción más popular fue la Raspberry Pi 3 Modelo B+ con un precio estimado de 35 USD, seguido por una Realtek Ameba RTL8195 con un precio estimado de 24 USD. Arduino One con un precio estimado de 23 USD compartiría la tercera posición con NodeMCU ESP8266 con un precio estimado de 5 USD. En general, la familia Arduino es la opción más popular para diseñar estaciones de monitoreo.

7.2.3 Resultados

Detalles, discusiones y resultados adicionales pueden obtenerse en [172]. Aquí se sintetizan las principales conclusiones relevantes al capítulo:

- No existe consenso sobre los métodos de calibración alrededor de los sensores de bajo costo basados en Internet de las cosas. Sin embargo, esto representa un tema esencial para la confiabilidad y exactitud del esquema de monitoreo.

- No es concluyente y permanece abierto el modo en que los marcos de medición se asocian con la semántica de los datos para asegurar la repetibilidad, consistencia, extensibilidad y comparación de los resultados a lo largo del proceso de medición.

- Existe una oportunidad de integración de las estrategias de recolección de datos respecto de los sistemas de recomendación.

- El empleo de Internet de las Cosas para el monitoreo de material particulado permitió incrementar el área de cobertura y densidad debido al balance entre precisión, confiabilidad, y costo. Ello es una oportunidad para aplicaciones tales como monitoreo de calidad el aire, monitoreo de pacientes, monitoreo de polvo o polución, entre otras.

- La coordinación distribuida y el procesamiento de datos constituyen un desafío desde el punto de vista de la cobertura, pero al mismo tiempo una oportunidad para desplegar sistemas de recolección baratos, distribuidos y escalables que puedan integrarse con sistemas de toma de decisiones en tiempo real.

- La limitación principal de las estrategias de monitoreo de material particulada basada en Internet de las Cosas es que ellas focalizan en obtener una medida (valor numérico), aunque no suele considerarse la relación entre el significado del dato y su análisis. El desafío reside en articular la heterogeneidad de los datos a modelos de toma de decisiones incrementales que interpreten su significado para proveer recomendaciones.

- La heterogeneidad subyacente de los sensores, métodos de medición (y su compatibilidad), y el significado de los datos es un desafío abierto para el desarrollo de una toma de decisiones consistente y los esquemas de recomendación.

- La utilización de ontologías de medición para identificar, relacionar y articular diferentes conceptos involucrados (discriminando apropiadamente el significado

de los datos) no ha sido detectado en los trabajos analizados. Esto constituye un desafío dada la heterogeneidad del contexto.

De este modo, el empleo de ontologías de medición para soportar el diseño experimental parece pendiente. Dada la factibilidad de emplear dispositivos basado en Internet de las Cosas para monitorear material particulado, el desafío reside en articular 1) La entidad bajo análisis; 2) Las características analizadas para la entidad; 3) La asociación entre categorías con un método de cuantificación; 4) El sensor/dispositivo a emplear para implementar un método para la característica bajo análisis; 5) Cómo cada medida se identifica, sigue, e interpreta; 6) Cómo la interpretación de la medida podría guiar recomendaciones.

7.3 Descripción del Escenario de Uso

Las enfermedades respiratorias son la tercera causa de muerte en el mundo y en Argentina. Tanto la morbilidad como la mortalidad están asociadas a los picos estacionales de enfermedades respiratorias transmisibles tipo influenza.

De acuerdo con la OMS, las infecciones de las vías respiratorias inferiores son la enfermedad transmisible más letal. El aerosol consiste en la suspensión de partículas sólidas o líquidas en el aire (al menos por segundos) y se dispersa mediante corriente de aires.

La aerosolización de diferentes grupos de virus y bacterias en condiciones naturales [174] y la persistencia de coronavirus (SARS-COV-1 y 2) en aerosoles han sido demostradas [175].

La Provincia de la Pampa es característica por sus vientos, llanuras, y región semi-árida lo que constituye un entorno propicio para la propagación de diferentes tipos de partículas mediante el aire.

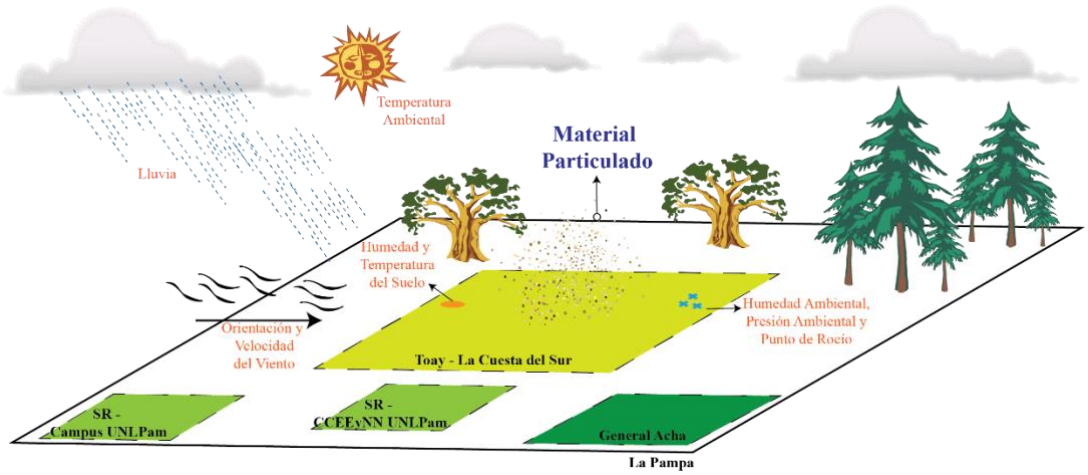

Figura 67 Escenario para el Monitoreo de Material Particulado

El objetivo del proyecto de medición consiste en monitorear el material particulado. Para ello, se define como entidad bajo análisis el *material particulado* (Ver Figura 67).

Sin embargo, diversos elementos del contexto deben considerarse respecto de la medición del material particulado además de la temperatura y humedad ambiental que afectan sus medidas. Es decir, deben considerarse aquí:

- Temperatura ambiental
- Presión Ambiental
- Humedad ambiental
- Lluvia
- Orientación del Viento
- Velocidad del Viento
- Humedad del Suelo
- Temperatura del Suelo

La combinación de temperatura, presión, y humedad ambiental inciden directamente sobre las condiciones del suelo y la posibilidad de partículas de arena suspendidas en el aire. Sin descartar el hecho de cómo ellas podrían incidir en el instrumento de medición de material particulado.

La lluvia, humedad del suelo y temperatura del suelo permiten contrastar el principal aportante de partículas y su condición comparativa respecto de la situación en el aire.

La orientación y velocidad del viento permite analizar la posibilidad en que los flujos de aire afecten las condiciones del al suelo (por ejemplo, humedad y temperatura) para propagar partículas en una cierta orientación.

El Instituto Nacional de Ciencias de la Tierra y Ambientales de La Pampa (INCITAP) cuenta con estaciones de monitoreo ambientales instaladas en el campus de la Universidad Nacional de La Pampa (UNLPam) localizado a 10 Km de la ciudad de Santa Rosa, La Facultad de Ciencias Exactas y Naturales en la zona urbana de Santa Rosa, y una estación de monitoreo instalada en la localidad de General Acha (La Pampa). Adicionalmente, se instaló una estación de monitoreo en la manzana 33 lote 9 del Club de Campo “La Cuesta del Sur” a 15 km de Santa Rosa (Localidad de Toay). De este modo, puede contrastarse el material particulado de la zona urbana de las localidades de Santa Rosa, General Acha y Toay.

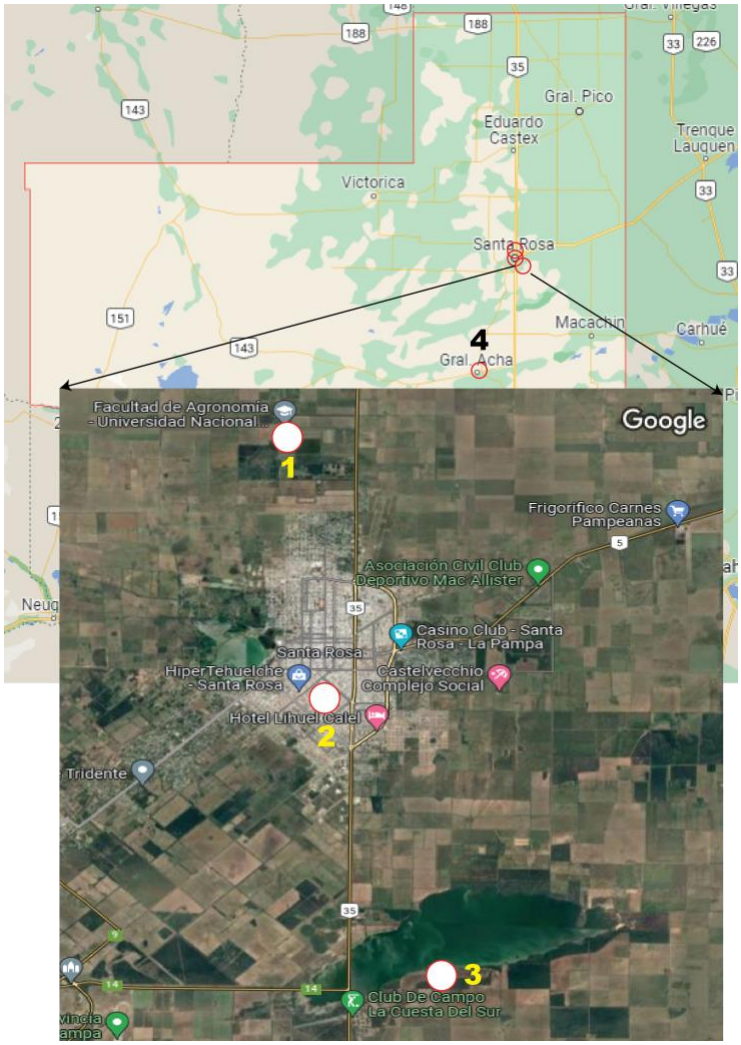

Figura 68 Mapa de Google con la Zona bajo monitoreo

La Figura 68 indica los puntos 1 a 3 en la ubicación próxima a la zona húmeda de La Pampa en la localidad de Santa Rosa. La zona inferior de la figura ubica los puntos en un mapa del satélite. Puede apreciarse el campus de la UNLPam como 1 en las afueras del ejido urbano, al igual que 3 localizado a la veda de la laguna Bajo Giuliani correspondiente al Club de Campo "La Cuesta del Sur". El número 2 señala la localización de la Facultad de Ciencias Exactas en la zona urbana de Santa Rosa. El número 4 indica la ubicación de la localidad de General Acha al sur oeste de Santa Rosa.

Tabla 35 Sensores Empleados para la Cuantificación de Atributos y Propiedades de Contexto

Sensor	Características	Atributos
DHT-22 (Precio: ~10 USD)	Rango Temperatura: -40º a 176º F (-40ºC a 80ºC) Rango Humedad: [0-100]%	Temperatura y Humedad Ambiental
BMP180 (Precio: ~5 USD)	Rango de Presión: 300-1100 hPa	Presión Ambiental y altitud

	(9000m a -500m Sobre el nivel del mar)	
RC-37 (Precio: ~7 USD)	Capacidad de carga del relé: 250V 10A (CA) 30V 10A (CC)	Lluvia (Detección)
Davis 6410 (Precio: ~150 USD)	Rango de la Velocidad de Viento: 1 a 89 m/s Resolución de Pantalla: 1º en pantalla numérica	Anemómetro (Velocidad del Viento) y dirección
Keyees – Humedad del Suelo (Precio: ~3 USD)	Rango Humedad: [0-100]%	Humedad del Suelo
Vegetronix THERM-200 (Precio: ~40 USD)	Rango Temperatura: -40º a 185º F (-40ºC a 85ºC)	Temperatura del Suelo
SDS011 (Precio: ~23 USD)	Rango de Medición: 0,0 a 999,9 $\mu g/m^3$	Material Particulado

La Tabla 35 describe los sensores empleados para medir cada uno de los atributos/propiedades de contexto descritos. Dichos sensores se calibran en una cámara en laboratorio bajo condiciones contraladas antes de su instalación. El proceso básico es incorporar el dispositivo con los sensores en patrones de referencia para los diferentes elementos y contrastarlos respecto de los valores arrojados por los sensores. A partir de allí, se monitoreo durante un día completa los valores y se estima el coeficiente o función de corrección por cada atributo (o propiedad de contexto).

Las estaciones de monitoreo de INCITAP suben los datos de las medidas al sitio global para el monitoreo de material particulado, tal y como puede apreciarse en la Figura 69. Como se puede apreciar, la gráfica para PM1 figura sin dato dado que el sensor de SDS011 es un sensor de PM2.5.

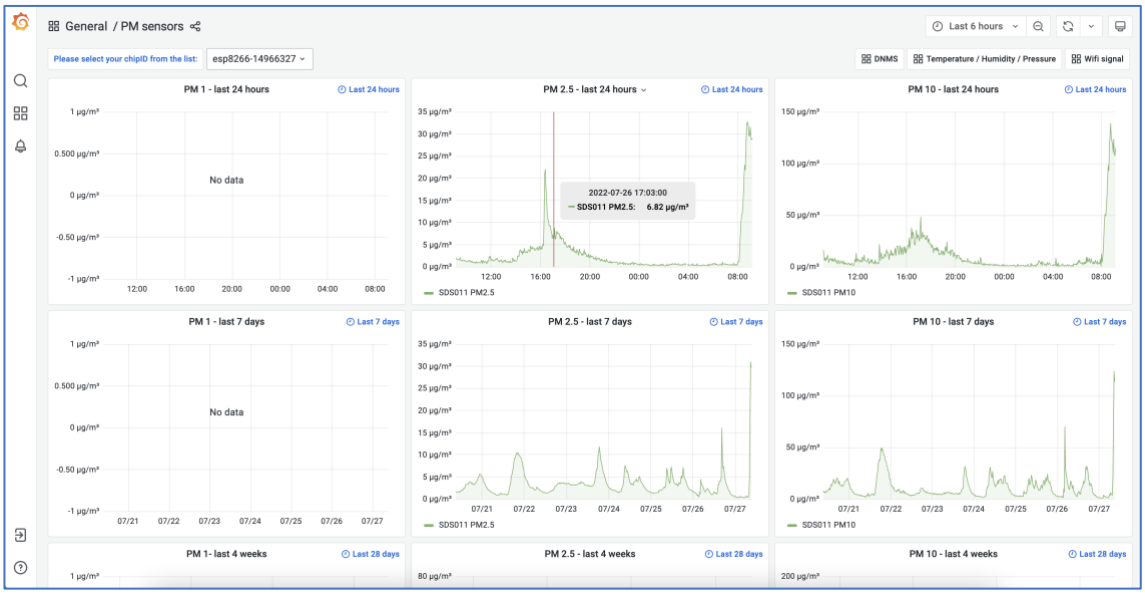

Figura 69 Datos de Material Particulado al 27 de Julio de 2022 - Estaciones de INCITAP

Los datos en formato de texto plano están disponibles mediante el repositorio de la comunidad de sensores global disponible en https://archive.sensor.community/.

7.4 Aplicación de la Distancia Compuesta

El Anexo A.1 describe una muestra de datos con intervalos de una hora para el 13 de abril de 2022 asociado con cada una de las estaciones de monitoreo introducidas en la Figura 68. Los mismos será utilizados como referencia para describir el cálculo de la distancia compuesta y la búsqueda de proyectos similares en caso de que no se cuente con experiencia previa. De hecho, la estación de monitoreo de La Cuesta (Toay) no cuenta con experiencia previa. Debido a ello, se detallará como a partir de los datos de ejemplo, dicha estación es capaz de localizar estaciones similares.

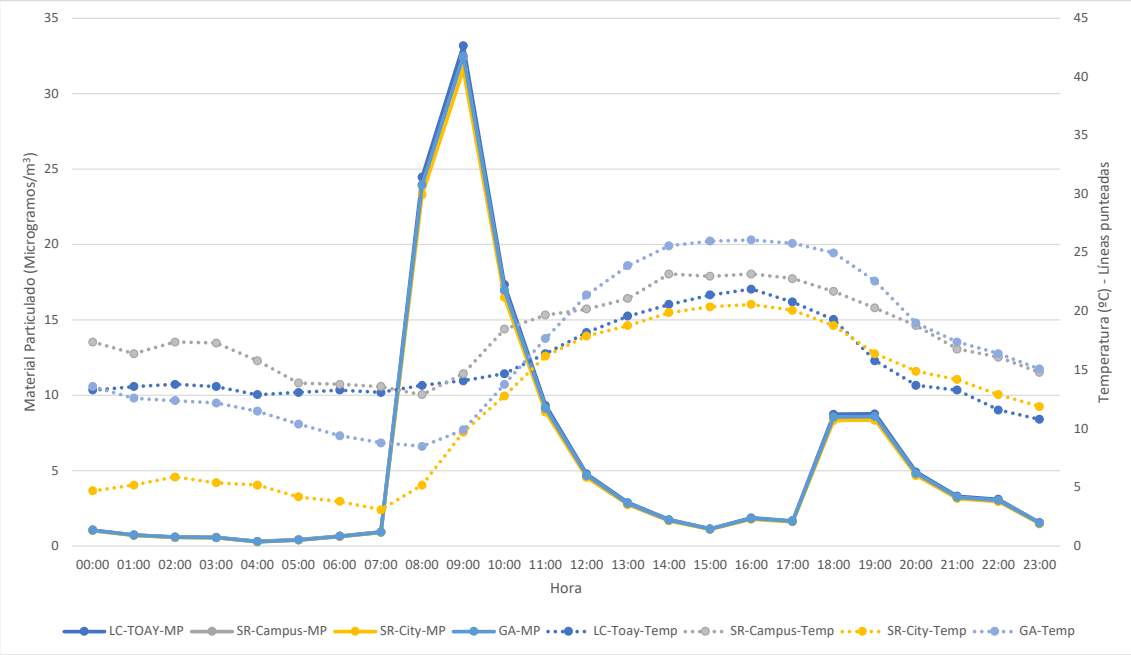

Figura 70 Evolución del Material Particulado y Temperatura Ambiental por Estación el 13-abr-2022

La Figura 70 expone el comportamiento de la temperatura Ambiental junto con las lecturas de material particulado para el 13 de abril de 2022 en los cuatro puntos de monitoreo. El material particulado se expresa en microgramos por metro cúbico respecto al eje ordenado izquierdo empleando líneas continuas, mientras que las temperaturas se expresan en grados Celsius respecto del eje ordenado derecho utilizando líneas punteadas.

Se han mantenido los mismos de colores para expresar material particulado y temperatura de la misma estación. Por ejemplo, la línea punteada en azul oscuro representa la temperatura para la estación de La Cuesta del Sur (Toay) y el mismo color con una línea punteada expresa las lecturas de material particulado.

Notar que las lecturas del material se comportan en forma similar, mostrando sus mayores lecturas entre 7:00 AM y 1:00 PM y entre 5:00 PM y 8:00 PM. Lo cual es de

esperar dado que son horarios asociados con la actividad típica de la zona (por ejemplo, trabajo, niños en la escuela, etc.).

La Figura 71 representa el diagrama de caja para la humedad ambiental de las estaciones bajo análisis. Como es posible observar se trata de un valor relativo que presenta un comportamiento diferente para el mismo día en las cuatro estaciones.

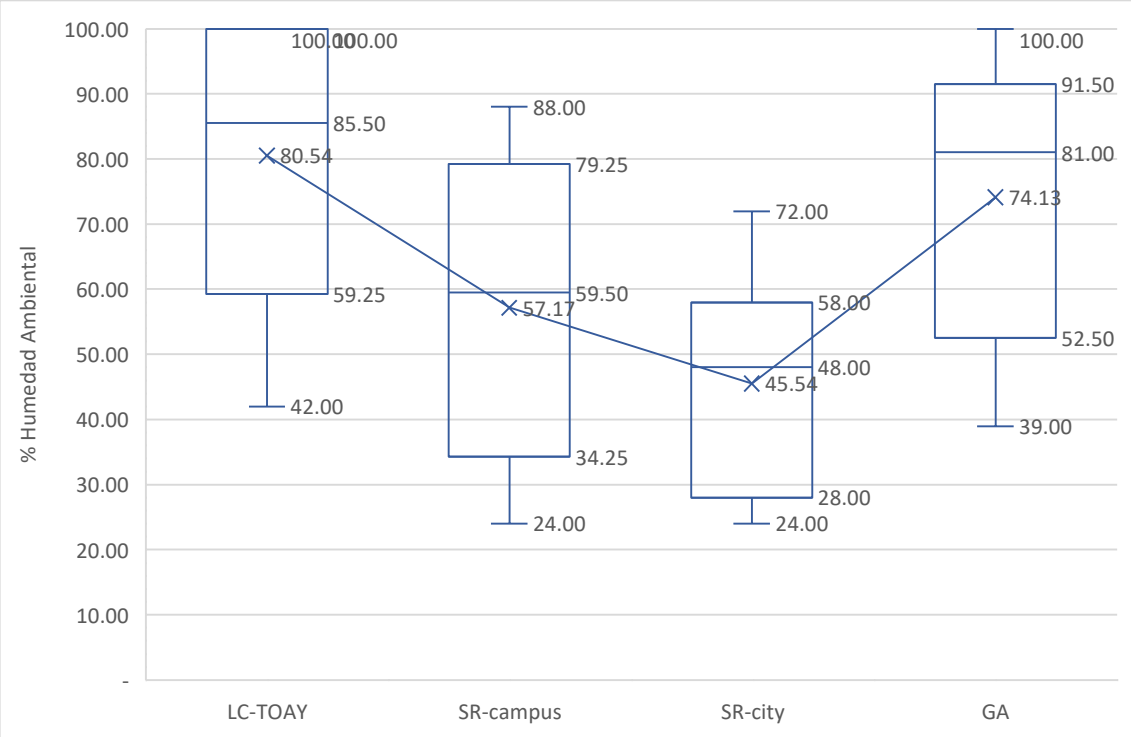

Figura 71 Boxplot para la Humedad Ambiental de las Estaciones de Monitoreo el 13-abr-2022

La estación de monitoreo de La Cuesta presenta mayor concentración de valores elevados con una mediana de 80,54%, una media de 85,50%. Ello se debe a que el barrio se encuentra en las orillas del Bajo Giuliani, lo cual influye en las medidas de humedad ambiente como en aquellas asociadas con la humedad del suelo.

La estación de General Acha tuvo una mediana de 74.13% con una media aritmética de 81%. En la ciudad de Santa Rosa la humedad promedio fue de 48% mientras que la mediana fue 45,54%. Estas dos últimas se asocian con zonas urbanizadas. Sin embargo, el campus de la UNLPam tuvo un promedio de 48% de humedad con una mediana de 45,54%. La diferencia entre mediana y mediana se deben a valores atípicos que jalan la media.

La Figura 72 muestra las tablas de correlaciones para el 13 de abril utilizando los datos del 13 de abril. Este cálculo al igual que las medidas descriptivas por métrica es llevado adelante por PAbMM a través de las funciones de análisis en la capa analítica, como así también por la estación de monitoreo mediante el cómputo incremental en el adaptador de medición introducido en la sección 6.2.1.

A los efectos del material particulado, existen asociaciones interesantes que surgen como, por ejemplo, en La Cuesta Toay habría vinculación del material particulado con la velocidad del viento y la presión ambiental. Por otro lado, en el campus de UNLPam las lecturas de material particulado parecerían asociarse con la temperatura del suelo y presión ambiental. Sin embargo, debe recordarse que se tienen datos de un solo día y que tal estimación se circunscribe a un momento específico, de allí la importancia del monitoreo y cálculo incremental.

Adicionalmente, puede observarse en La Cuesta (Matriz de correlación LC-Toay enFigura 72) relaciones entre la temperatura ambiental (*envtemp*) y la humedad ambiental (*envhum*); temperatura ambiental (*envtemp*) y temperatura del suelo (*soiltemp*); humedad ambiental (*envhum*) y temperatura del suelo (*soiltemp*).

LC-Toay

	pm (ug/m3)	*envtemp*	*envhum*	*envpress*	*rain*	*windspeed (k*	*windorientat*	*soilmoisture*	*soiltemp*
pm (ug/m3)	1.00								
envtemp	(0.08)	1.00							
envhum	0.21	(0.88)	1.00						
envpress	0.55	0.01	(0.12)	1.00					
rain	(0.14)	(0.17)	0.19	(0.30)	1.00				
windspeed (km/h)	0.64	0.23	(0.03)	0.69	(0.16)	1.00			
windorientation	(0.29)	(0.19)	0.06	(0.10)	0.22	(0.36)	1.00		
soilmoisture	0.00	0.00	0.00	(0.00)	(0.00)	0.00	-	1.00	
soiltemp	(0.27)	0.79	(0.91)	(0.10)	(0.13)	(0.23)	0.03	0.00	1.00

SR-Campus

	pm (ug/m3)	*envtemp*	*envhum*	*envpress*	*rain*	*windspeed (k*	*windorientat*	*soilmoisture*	*soiltemp*
pm (ug/m3)	1.00								
envtemp	(0.21)	1.00							
envhum	0.02	(0.92)	1.00						
envpress	0.52	0.19	(0.51)	1.00					
rain					1.00				
windspeed (km/h)	(0.08)	0.79	(0.79)	0.39		1.00			
windorientation	(0.43)	(0.26)	0.49	(0.60)		(0.10)	1.00		
soilmoisture	(0.00)	0.00	(0.00)	(0.00)		(0.00)	-	1.00	
soiltemp	(0.55)	0.50	(0.22)	(0.61)		0.03	0.11	(0.00)	1.00

SR-City

	pm (ug/m3)	*envtemp*	*envhum*	*envpress*	*rain*	*windspeed (k*	*windorientat*	*soilmoisture*	*soiltemp*
pm (ug/m3)	1.00								
envtemp	(0.01)	1.00							
envhum	(0.04)	(0.94)	1.00						
envpress	0.10	(0.79)	0.57	1.00					
rain					1.00				
windspeed (km/h)	0.01	0.95	(0.94)	(0.63)		1.00			
windorientation	0.40	0.08	0.02	(0.16)		0.10	1.00		
soilmoisture	(0.00)	(0.00)	(0.00)	(0.00)		(0.00)	-	1.00	
soiltemp	(0.41)	0.75	(0.62)	(0.82)		0.60	(0.18)	0.00	1.00

GA

	pm (ug/m3)	*envtemp*	*envhum*	*envpress*	*rain*	*windspeed (k*	*windorientat*	*soilmoisture*	*soiltemp*
pm (ug/m3)	1.00								
envtemp	(0.24)	1.00							
envhum	0.26	(0.98)	1.00						
envpress	0.46	(0.50)	0.43	1.00					
rain					1.00				
windspeed (km/h)	(0.18)	0.77	(0.79)	(0.46)		1.00			
windorientation	0.19	0.63	(0.61)	(0.06)		0.46	1.00		
soilmoisture	0.00	0.00	-	0.00		0.00	-	1.00	
soiltemp	(0.38)	0.93	(0.88)	(0.75)		0.71	0.46	0.00	1.00

Figura 72 Matriz de Correlación para las Estaciones de Monitoreo (Datos del 13 de abril de 2022)

Para el campus de la UNLPam (Matriz de correlación SR-Campus en Figura 72) las principales asociaciones serían entre temperatura ambiental y humedad ambiente; presión ambiental y velocidad del viento; temperatura ambiental y velocidad del viento; orientación del viento y presión ambiental; presión ambiental y temperatura del suelo.

Para la ciudad de Santa Rosa (Matriz de correlación SR-Campus en Figura 72) las asociaciones principales estarían entre temperatura y humedad ambientales; temperatura ambiental y presión; velocidad del viento respecto a temperatura y humedad ambiental; temperatura del suelo respecto a la temperatura, humedad, presión ambiental, y velocidad del viento.

Independientemente de ello, notar que, aunque las asociaciones parecen lógicas en muchas matrices, no se replican de igual modo en cada sitio. Por ejemplo, la asociación entre material particulado y velocidad del viento que se expone en La Cuesta suena lógico, aunque no se replica en las restantes estaciones. Ello no indica que esa asociación no existe o es errónea, sino que los datos de las medidas se comportan en forma diferente. Ese es uno de los motivos por el cual la distancia compuesta incorpora la perspectiva comportamental.

Dado que La Cuesta es la última estación incorporada, se detalla a continuación el modo en que PAbMM calcularía la distancia compuesta usando los mismos datos del anexo a.1 para ordenar las similitudes de proyectos y ordenar el camino de búsqueda de recomendaciones previo a ir a la memoria organizacional.

Se toman los datos del apéndice a.1, se discretizan en 5 intervalos igualmente espaciados tomando como referencia mínimo y máximo para su definición. Luego, se distribuyen los valores medidos en los intervalos y se contabiliza la ocurrencia por intervalo. Ello permite estimar una probabilidad empírica a partir de la frecuencia de valores por intervalo para cada métrica. A partir de ellos se calcula la Ecuación 9 (Distancia Comportamental para un Atributo) para cada una de las métricas, sean propiedades de contexto o atributo. Dado que se utiliza la misma definición de proyecto en todas las estaciones de monitoreo, la estructura de los proyectos es idéntica.

Tabla 36 Cálculo de la distancia comportamental para material particulado (Ecuación 9)

	LC-TOAY	SR-campus	SR-city	GA
LC-Toay	1.00	0.89	0.92	0.84
SR-Campus	0.89	1.00	0.92	0.81
SR-City	0.92	0.92	1.00	0.84
GA	0.84	0.81	0.84	1.00

La Tabla 36 describe la distancia comportamental para el material particulado tomando como entidad el sitio de monitoreo (es decir, La Cuesta -LC-Toay-, Campus de la UNLPam -SR-campus-, Ciudad de Santa Rosa -SR-city-, y General Acha -GA-). Se reitera su cálculo para cada una de las propiedades contextuales involucradas en el caso de estudio.

Tabla 37 Cálculo de la distancia comportamental para la temperatura ambiental (Ecuación 9)

	LC-TOAY	SR-campus	SR-city	GA
LC-Toay	1.00	0.70	0.50	0.63
SR-Campus	0.70	1.00	0.41	0.57
SR-City	0.50	0.41	1.00	0.39
GA	0.63	0.57	0.39	1.00

La Tabla 37 sintetiza la distancia comportamental para la temperatura ambiental por estación de monitoreo.

Tabla 38 Cálculo de la distancia comportamental para la temperatura del suelo (Ecuación 9)

	LC-TOAY	SR-campus	SR-city	GA
LC-Toay	1.00	0.09	0.57	0.68
SR-Campus	0.09	1.00	0.07	0.09
SR-City	0.57	0.07	1.00	0.79
GA	0.68	0.09	0.79	1.00

La Tabla 38 indica la distancia comportamental para la temperatura del suelo por estación de monitoreo siguiendo la aplicación de la Ecuación 9.

Tabla 39 Cálculo de la distancia comportamental para la humedad ambiental (Ecuación 9)

	LC-TOAY	SR-campus	SR-city	GA
LC-Toay	1.00	0.54	0.29	0.83
SR-Campus	0.54	1.00	0.64	0.64
SR-City	0.29	0.64	1.00	0.39
GA	0.83	0.64	0.39	1.00

La Tabla 39 describe la distancia comportamental para la humedad ambiental por estación de monitoreo. Notar que la zona pintada con gris señala la diagonal principal comparando la estación consigo misma, mientras que la matriz triangular superior es análoga a la inferior.

Tabla 40 Cálculo de la distancia comportamental para la humedad del suelo (Ecuación 9)

	LC-TOAY	SR-campus	SR-city	GA
LC-Toay	1.00	1.00	1.00	1.00
SR-Campus	1.00	1.00	1.00	1.00
SR-City	1.00	1.00	1.00	1.00
GA	1.00	1.00	1.00	1.00

La Tabla 40 sintetiza el cálculo de la distancia comportamental para la humedad del suelo.

Tabla 41 Cálculo de la distancia comportamental para la presión ambiental (Ecuación 9)

	LC-TOAY	SR-campus	SR-city	GA
LC-Toay	1.00	0.00	0.00	0.70
SR-Campus	0.00	1.00	0.00	0.00
SR-City	0.00	0.00	1.00	0.12
GA	0.70	0.00	0.12	1.00

La Tabla 41 resume los resultados obtenidos para la distancia comportamental para la presión ambiental.

Tabla 42 Cálculo de la distancia comportamental para la lluvia registrada (Ecuación 9)

	LC-TOAY	SR-campus	SR-city	GA
LC-Toay	1.00	0.85	0.85	0.85
SR-Campus	0.85	1.00	1.00	1.00
SR-City	0.85	1.00	1.00	1.00
GA	0.85	1.00	1.00	1.00

La Tabla 42 sintetiza los resultados de la distancia comportamental para la lluvia registrada por estación de monitoreo.

Tabla 43 Cálculo de la distancia comportamental para la velocidad del viento (Ecuación 9)

	LC-TOAY	SR-campus	SR-city	GA
LC-Toay	1.00	0.72	0.59	0.94
SR-Campus	0.72	1.00	0.73	0.70
SR-City	0.59	0.73	1.00	0.58
GA	0.94	0.70	0.58	1.00

La Tabla 43 presenta los resultados de la distancia comportamental para la velocidad del viento.

Tabla 44 Cálculo de la distancia comportamental para la orientación del viento (Ecuación 9)

	LC-TOAY	SR-campus	SR-city	GA
LC-Toay	1.00	0.42	0.56	0.33
SR-Campus	0.42	1.00	0.53	0.54
SR-City	0.56	0.53	1.00	0.19
GA	0.33	0.54	0.19	1.00

La Tabla 44 describe los resultados de la distancia comportamental para la orientación del viento entre estaciones de monitoreo. Una vez calculadas las distancias comportamentales por atributo, se calcula la distancia comportamental interna para la entidad.

Tabla 45 Cálculo de la distancia comportamental interna para las entidades (idistbeh, Ecuación 10)

Entidades	LC-TOAY	SR-campus	SR-city	GA
LC-Toay	-	0.11	0.08	0.16
SR-Campus	0.11	-	0.08	0.19
SR-City	0.08	0.08	-	0.16
GA	0.16	0.19	0.16	-

La Tabla 45 sintetiza el cálculo de la distancia comportamental interna tomando como único atributo el material particulado. En ella, cuanto más próximo a cero los valores más similares son las entidades, y cuanto más próximo a 1 mayor la diferencia.

Tabla 46 Cálculo de la distancia comportamental externa para los contextos (edistbeh, Ecuación 13)

Contextos	LC-TOAY	SR-campus	SR-city	GA
LC-Toay	-	0.46	0.45	0.25
SR-Campus	0.46	-	0.45	0.43
SR-City	0.45	0.45	-	0.44
GA	0.25	0.43	0.44	-

La Tabla 46 sintetiza la distancia comportamental externa para los contextos tomando como propiedades de contexto la temperatura ambiental, la temperatura del suelo, humedad ambiental, humedad del suelo, velocidad del viento, dirección del viento, lluvia, y presión ambiental (es decir, 8 propiedades de contexto en común).

Tabla 47 Cálculo de la distancia interna según la Ecuación 14

idist	LC-TOAY	SR-campus	SR-city	GA
LC-Toay	-	0.06	0.04	0.08
SR-Campus	0.06	-	0.04	0.10
SR-City	0.04	0.04	-	0.08
GA	0.08	0.10	0.08	-

La Tabla 47 describe la distancia interna considerando idéntica similitud estructural.

Tabla 48 Cálculo de la distancia externa según la ecuación 15

edist	LC-TOAY	SR-campus	SR-city	GA
LC-Toay	-	0.230	0.227	0.126
SR-Campus	0.230	-	0.226	0.216
SR-City	0.227	0.226	-	0.221
GA	0.126	0.216	0.221	-

La Tabla 48 describe la distancia externa considerando idéntica similitud estructural.

Tabla 49 Distancia compuesta según la ecuación 16

cdist	LC-TOAY	SR-campus	SR-city	GA
LC-Toay	-	0.1425	0.1342	0.1023
SR-Campus	0.1425	-	0.1342	0.1560
SR-City	0.1342	0.1342	-	0.1494
GA	0.1023	0.1560	0.1494	-

La Tabla 49 describe el cómputo de la distancia compuesta a partir de la distancia interna y externa, contemplando igual peso para las mismas. De este modo, dada la estación de monitoreo de La Cuesta, si PAbMM debiera buscar recomendaciones en proyectos similares recorrería los restantes en el siguiente orden: 1) General Acha (0.1023), 2) Ciudad de Santa Rosa (0.1342), y 3) Campus de la UNLPam en Santa Rosa (0.1425). Claro que estos resultados se actualizan en forma permanente ante el arribo de nuevas medidas, por lo que se actualizan incrementalmente en consecuencia y el orden a seguir dependerá del momento en que deban accederse a las recomendaciones complementarias.

Esto implica una colaboración entre adaptadores de medición para informar datos pertinentes respecto de PAbMM introducida en el capítulo 6 con los detectores de cambios, estimaciones incrementales, registros de integridad y transmisiones indirectas a partir del registro distribuido. Es decir, el orden de búsqueda depende de la calidad del dato recibido y allí es donde el adaptador de medición aporta su valor añadido, informando aquellos con suficiente sustento. En este ejemplo, se simplificó a datos por hora para describir los principales cálculos, no obstante, en tiempo real implicaría numerosas medidas informadas para el mismo intervalo de tiempo. Así, el adaptador informa aquello que estadísticamente tiene sustento para producir un cambio en la serie de datos que es aprovechado por la distancia compuesta para calcular las actualizaciones en consecuencia.

7.5 Conclusiones Generales del Capítulo

Este capítulo introdujo el rol del material particulado y la importancia de su monitoreo permanente en la salud de las personas. Un aspecto esencial es que, dado el tamaño asociado con el material particulado, este podría invadir diferentes órganos del cuerpo produciendo diferentes enfermedades, afectando el tracto respiratorio, o influenciando respecto de enfermedades prexistentes [161]–[163]. Este constituye uno de los principales motivantes del Desarrollo de diversas estrategias para estudiar y monitorear la calidad del aire, analizando el material particulado en concentración y composición [164]–[166].

Se sintetizaron las estrategias actuales empleadas para monitorear el material particulado, equipamientos y sensores típicos, así como los principales desafíos asociados. Entre los presentes desafíos, puede mencionarse que:

- No es concluyente y permanece abierto el modo en que los marcos de medición se asocian con la semántica de los datos para asegurar la repetibilidad, consistencia, extensibilidad y comparación de los resultados a lo largo del proceso de medición.

- La limitación principal de las estrategias de monitoreo de material particulada basada en Internet de las Cosas es que ellas focalizan en obtener una medida (valor numérico), aunque no suele considerarse la relación entre el significado del dato y su análisis. El desafío reside en articular la heterogeneidad de los datos a modelos de toma de decisiones incrementales que interpreten su significado para proveer recomendaciones.

A tales efectos, se planteó como escenario de uso el monitoreo de material particulado en cuatro puntos diferentes de La Provincia de La Pampa. El campus de la UNLPam (en las afueras de Santa Rosa), la ciudad de Santa Rosa (ejido urbano), La Cuesta del Sur (Toay) a 15km de la ciudad de Santa Rosa, y General Acha. Este estudio originalmente iniciado por INCITAP, permitió abordar una problemática real en una zona con clima semiárido, frecuencia de vientos y con velocidades importantes sobre una llanura que posibilita la propagación de material particulado.

Para ejemplificar el aporte al esquema de procesamiento desde el punto de vista lógico, tanto los adaptadores de medición como PAbMM interpretan la importancia de atributos y propiedades contextuales mediante la ponderación asociada. Se estiman incrementalmente las medidas estadísticas a medida que arriban datos. Para ejemplificar parte de los cálculos que se realizan, a) se tomó como día de referencia el 13 de abril de 2022, b) se sintetizó la correlación de variables (limitado a los datos del anexo a.1) como se introdujo en el capítulo 6, c) Se detallaron los cálculos para las ecuaciones introducidas en el capítulo 4 para cuantificar las distancias, d) Se obtiene la priorización de proyectos basado en comportamiento y estructura para articularse con

PAbMM (Ver capítulo 5) para organizar la búsqueda de recomendaciones si fuere necesario.

Como se pudo apreciar, la distancia compuesta constituye un concepto dinámico que va cambiando con el arribo de nuevas medidas y se emplea en el momento preciso en que debe determinarse el orden de consulta. Esto implica una colaboración entre adaptadores de medición para informar datos pertinentes respecto de PAbMM introducida en el capítulo 6 con los detectores de cambios, estimaciones incrementales, registros de integridad y transmisiones indirectas a partir del registro distribuido. De este modo, la distancia compuesta se actualiza en forma incremental con las medidas, lo que produce cambios continuos en la prioridad de proyectos para buscar recomendaciones. Este cálculo de distancia es incremental ante el arribo de las medidas y no forma parte del costo de un plan de consulta sobre la memoria organizacional. De hecho, la distancia permite avanzar en la búsqueda de recomendaciones adicionales (cuando se requiera), focalizando en los proyectos *actualmente* más similares.

Así, la distancia compuesta no requiere emitir consulta a la memoria organizacional (MO) de PAbMM para analizar similitudes de proyectos. Por el contrario, se actualiza dinámicamente y en tiempo real para determinar el orden en que los proyectos de la MO se visitarán para buscar recomendaciones adicionales cuando sea necesario. Dicho orden o priorización responde a lo que exactamente está sucediendo en el momento de consulta con las entidades bajo análisis.

Capítulo 8

Conclusiones

Capítulo 8 - Conclusiones

Los capítulos han desarrollado la temática basada en las contribuciones documentadas logradas durante el proceso de la presente investigación en un todo de acuerdo con el objetivo general, el cual indica:

"Desarrollar una estrategia de recomendación en memoria, basado en entidades bajo monitoreo semánticamente similares, para mejorar la precisión y reutilización de conocimiento y/o experiencia previa ante situaciones nuevas y-o no tipificadas, en los cuales una decisión requiera de cursos de acción como soporte".

A partir del mismo, se desprenden los siguientes objetivos específicos:

1) Mejorar la precisión en las recomendaciones ante el tomador de situaciones para una situación dada.

2) Reutilizar el conocimiento y/o experiencia previa ante situaciones nuevas y-o no tipificadas, basado en similitud semántica de las entidades bajo monitoreo.

3) Detectar definiciones contradictorias que pudieren afectar las recomendaciones asociadas a una decisión.

4) Detectar atributos homónimos y evitar la redundancia entre atributos en sus definiciones.

5) Acotar el espacio de búsqueda en el repositorio columnar con grandes volúmenes de datos a partir de la similitud semántica de entidades bajo monitoreo.

Principales Contribuciones de la Tesis

La Tabla 50 provee una síntesis de las principales contribuciones respecto de los objetivos específicos y del objetivo general. Esto permite guiar el desarrollo de las conclusiones y detallar el modo en que cada contribución aportó al objetivo de la tesis.

Tabla 50 Síntesis de las Principales Contribuciones Respecto de los Objetivos Específicos

Contribuciones	Fuente	(1) Precisión de las Recomenda-ciones	(2) Reutilizar conocimiento y-o experiencia	(3) Detectar definiciones contradict-orias	(4) Evitar redundancia de atributos	(5) Acotar el espacio de búsqueda
ECINCAMI	3.1	X	X	X	X	
BriefPD	3.3	X	X	X	X	

Estados de Entidad y Escenarios	3.1 4.3 5.3	X	X	X	X	
Similitud Estructural	4.1	X	X		X	
Similitud Comportamental	4.2	X	X	X	X	
Distancia Compuesta	4.4	X	X			X
CINCAMIMIS v2	5.2.1 5.4	X				
Brief	5.2.1 5.4	X				
Transmisión Indirecta de Medidas	5.2.2 5.2.3	X	X		X	
Correlaciones, Media y Desviaciones Incrementales de las medidas	5.5 6.1.1	X	X			X
Detectores de Cambio en línea basado en medidas	6.1.1 6.1.2 6.1.3	X				X
Descarte Selectivo	6.2	X				
Registro de Integridad mediante Merkle Tree	6.3	X	X		X	
Registro Distribuido basado en Cadena de Bloques	6.4				X	

Recordemos que C-INCAMI constituye un marco de referencia basada en una ontología de medición, por lo tanto, la extensión de la ontología originando ECINCAMI (Ver sección 3.1) permitió lo siguiente:

- Incorporar la posibilidad de gestionar medidas deterministas y estimativas asociadas con los atributos y propiedades de contexto.

- Gestionar datos complementarios a las medidas informadas, permitiendo transmitir una secuencia de video, audio, texto, datos geográficos, o una imagen como complemento.

- Se incorporó la figura de la fuente de datos y del adaptador de medición en el marco para aplicar restricciones a nivel de propiedades para facilitar el emparejamiento entre fuente y atributo,

- El concepto de grupos de seguimiento permite reunir una serie de fuentes de datos monitoreando un grupo de conceptos desde diferentes adaptadores de medición. Esto es útil porque permite analizar el concepto monitoreado desde diferentes puntos de vistas.

- La incorporación de estados de entidad permite definir estados observables del concepto bajo monitoreo a partir de sus atributos característicos. De este modo, el modelo de transición de estados asociados permite estimar la posibilidad de transitar entre estados de entidad para soportar la toma de decisiones.

- Análogamente a los estados de entidad, los escenarios permiten definir estados observables del contexto (o ambiente) caracterizado a través de sus propiedades de contexto. Así, su modelo de transición asociado permite estimar la posibilidad de transitar entre escenarios para soportar el proceso de toma de decisiones.

- El concepto de indicador fue extendido incorporando el rol de los escenarios y estados de entidad, para permitir interpretar una o más métricas basado en el escenario y estado actual de entidad. Tanto el escenario como el estado actual se estiman incrementalmente a medida que arriban los datos sin representar cargas significativas en el procesamiento. Esto permite incrementar la precisión en la toma de decisión, dado que la medida se interpreta circunscripta al escenario y estado actual, lo que focaliza en recomendaciones/cursos de acción conscientes del contexto y estado.

- La extensión de elementos para describir el proyecto de medición permite reusar conocimiento de contextos y entidades de otros proyectos con mayor precisión a la hora de comparar su similitud estructural y comportamental. La posibilidad de analizar el comportamiento en las entidades permite discernir entre atributos para diferentes entidades (incluso cuando ellos tuvieren idéntica definición como se expuso en el escenario de uso).

Ahora bien, la forma de intercambiar las definiciones de proyecto basado en ECINCAMI es esencial para mejorar la precisión en las recomendaciones. Esto es, si es posible definir el concepto de estados, escenarios, indicadores e interpretaciones basados en ellos, sería de esperar que un adaptador de medición pueda procesarlo. Allí es donde BriefPD permite incorporar una contribución importante (Ver sección 3.3). Este nuevo formato de intercambio de definiciones de proyecto de medición contribuye directamente a lo expresado por ECINCAMI porque permite la movilidad de tales conceptos a los efectos de implementar el proceso de medición. De este modo, se tiene que:

- Se incorporó una nueva estrategia para organizar el proyecto de monitoreo basado en la ontología desde la perspectiva del proyecto medición en lugar de los dispositivos. Esto representa una visión transversal, útil para relacionar y reutilizar los dispositivos entre diferentes proyectos. Cada proyecto se vería como una capa lógica, mientras que el conjunto de dispositivos se observaría como la capa física. El proyecto define en qué modo agrupar los dispositivos para compartirlos o reutilizarlo.

- El formato BriePD surge para soportar la estrategia de organización de proyecto de medición. El proyecto es sintácticamente interoperable porque el formato de datos promueve un entendimiento común entre dispositivos heterogéneos. A su vez, es semánticamente interoperable dado que el formato de datos se encuentra basado en la ontología de medición ECINCAMI. Así, puede indicarse que BriefPD es consistente, autocontenido, y libre de etiquetas. Consistente debido a que cada elemento en la definición se articula con un concepto de la ontología de medición. Es autocontenido porque todos los elementos del proyecto de medición se definen en un único mensaje. Finalmente, es libre de etiquetas dado que su contenido se especifica siguiendo la organización jerárquica de los conceptos en el modelo de navegación derivado de ECINCAMI.

- A partir de la implementación de referencia, BriefPD puede describir un proyecto de medición consumiendo cerca del 20% de su tamaño equivalente en JSON. Más aún, BriefPD puede ser convertido hacia y desde los formatos XML y JSON.

- BriefPD permite actualizaciones parciales a partir de la jerarquía del modelo de navegación derivado de ECINCAMI. De este modo, es posible reemplazar, remover, o actualizar alguna parte de la jerarquía asociada al proyecto de medición.

- BriefPD soporta la verificación parcial a partir de la lógica del árbol de Merkle. Esto permite contrastar total o parcialmente la definición de un proyecto de medición para evaluar similitudes o diferencias.

- Se proveyó tanto una implementación de referencia como su simulación asociada para demostrar la aplicación de este. La cápsula ocean con su código se entra disponible en https://codeocean.com/capsule/4500824/tree/v1, bajo los términos de la licencia Apache 2.0.

Además de los mencionado de los estados de entidad y escenarios (Ver secciones 3.1, 4.3, y 5.3), se incorporó un mecanismo para estimar las probabilidades empíricas en tiempo real a partir del arribo de las medidas. De este modo, se tiene que:

- Se incorporó el cómputo incremental de probabilidades empíricas de escenarios y estados de entidad a partir del arribo de las medidas informadas mediante el esquema de intercambio de medidas (alineado con la definición de proyecto),

- Se implementó la estimación de la probabilidad condicional entre estados de entidad y escenarios a partir de la determinación en línea de los estados y escenarios actuales.

- Esto importante para el indicador, por cuanto 1) Le permite conocer la posibilidad de ocurrencia en el proyecto y para la situación actual de los estados y escenarios (independientemente de la definición teórica); 2) Le permite estimar las probabilidades condicionales entre estados y escenarios lo cual no está contemplado en el modelo de transición de la definición; 3) La interpretación de la medida es actual respecto del estado, escenario y sus estimaciones asociadas.

Esto complementa los mismos aportes mencionados de ECINCAMI y BriefPD.

La similitud estructural se actualiza de acuerdo con los conceptos involucrados en ECINCAMI y considerando el esquema de intercambio de proyectos dato por BriefPD (Ver sección 4.1). A partir de ello, se estima la similitud estructural considerando:

- Entidades, atributos característicos y los estados de entidad asociados,
- Contextos, propiedades de contextos, y escenarios definidos.

Ahora bien, dado el escenario de uso del material particulado, se observó que idénticos proyectos replicados en zonas diferentes podrían presentar comportamientos

diferentes. Por tal motivo, se incorpora la similitud comportamental sobre los atributos y propiedades de contexto que caracterizan las entidades y escenarios respectivamente (Ver sección 4.2). De este modo, se combinan ambos conceptos en la distancia compuesta (Ver sección 4.3) la cual:

- Se implementa a partir de los metadatos de definición de proyecto (BriefPD), por lo que es calculable incluso a nivel de adaptador de medición.

- Puede discriminar entre similitud interna (dada por las entidades y sus estados) y similitud externa (basado por el contexto y sus escenarios).

- Permite proveer una lista basada en similitud de un conjunto de proyectos de medición para guiar la estrategia de búsqueda de recomendaciones sin efectuar una consulta previa. Los índices se calculan en memoria y se actualizan con el arribo de cada dato.

- Se provee una implementación de referencia en JAVA dentro de la librería composedIndex, disponible bajo los términos de la licencia Apache 2.0.

La actualización del esquema de intercambio de mediciones (Ver secciones 5.2.1 y 5.4) permitió incorporar las extensiones de la ontología ECINCAMI y soportar el intercambio de medidas basado en BriefPD. Esto permitió el intercambio de medidas estimativas, entre otros aspectos, lo que permite incrementar la precisión de los datos empleados por los indicadores para decidir sobre los cursos de acción y-o recomendaciones asociadas.

Además, la incorporación del formato de intercambio Brief (Ver sección 5.2.1 y 5.4) permitió lograr los siguientes aspectos:

- Informar flujos de medidas sin etiquetas y basado en el mapa de navegación de la definición de proyecto de medición derivado de ECINCAMI. Esto optimiza el tamaño de cada mensaje desde los adaptadores de medición con el consiguiente efecto positivo en equipos de recursos limitados (por ejemplo, ahorro de batería).

- Se complementa con la definición del proyecto para incorporar una única huella MD5 calculada a partir del contenido a transmitir y la especificación formal del proyecto. De este modo, es posible verificar en una comparación si los datos vienen inalterados para la misma definición de proyecto. Caso contrario, la huella diferirá, sea por el lado de la definición o las medidas a informar.

Siguiendo la definición de los proyectos de medición como guía, se definió un mecanismo para la transmisión indirecta de medidas estableciendo límites de vida para las medidas (Ver la sección 5.2.2 y 5.2.3). Así, un adaptador de medición que pierde conectividad podría seguir transmitiendo a partir de otro adaptador de medición, evitando la pérdida del dato sí y solo sí éste no excede un tiempo máximo de vida dado. Este tiempo máximo de vida garantiza que la capa de recolección de datos reciba datos relativamente recientes y evite ser sobrecargada con datos históricos. Por otro lado, es un modo de mitigar el riesgo de ausencia de datos desde el adaptador de medición, aunque no lo elimina dado que eventualmente una transmisión indirecta podría expirar antes de alcanzar la capa de recolección en la nube.

Gracias al empleo de la definición de proyecto mediante BriefPD y el intercambio de medidas a través de Brief, se ha podido lograr un cálculo incremental de media, desviación, y correlaciones (Ver secciones 5.5 y 6.1.1). Este cálculo es posible no solo a nivel de las capas centrales sino también a nivel de adaptador de medición. Así, cada uno puede actualizar las estimaciones ante el arribo de nuevos datos, pudiendo estimar la magnitud de las desviaciones respecto de las métricas procesadas. Esto se articula con los detectores de cambio y barreras temporales ya que permite informar medidas cuando se tiene suficiente prueba estadística de que un cambio en la serie de datos ha ocurrido. Esta estrategia puede complementarse con barreras temporales para dar prueba de vida. Las transmisiones basadas en detectores de cambio consumirían alrededor del 26% de todas las requeridas incluyendo las barreras temporales, lo cual es importante desde el punto de vista del ahorro de recursos, pero también desde la precisión. Esto último fundado en que los datos informados no son mera secuencia de valores sino una secuencia tal capaz de producir un cambio en la serie de datos para una métrica dada.

Se incorporaron técnicas de descarte selectivo para retener las métricas priorizadas en la definición del proyecto de medición (BriefPD) ante situaciones límite de consumo de recursos (Ver sección 6.2). El punto esencial de ello reside en garantizar que determinadas métricas (por ejemplo, la concentración del material particulado en el escenario de uso) siempre sean informadas aun cuando deban sacrificarse otras. Se evaluó esta aplicación junto con las barreras temporales y detectores de cambio, dando como resultado que los detectores de cambio eran una excelente oportunidad para racionalizar los datos informados y focalizar sobre datos que sustenten algún tipo de cambio en las métricas sin implicar descarte necesariamente.

Se incorporó un registro de integridad a nivel de adaptador de medición como de capa de recolección basado en árbol de Merkle (Ver Sección 6.3), sin que ello implique una sobrecarga de procesamiento. Esto permite mantener un seguimiento de la huella de las ventanas de medidas transmitidas, liberando recursos en el adaptador de medición, y ofreciendo un verificador de segundo factor para la capa de recolección. De este modo, a través de la comparación de huellas entre el adaptador y la capa de

recolección, es posible saber si las últimas ´n´ transacciones coinciden o no. Esto afecta directamente a la calidad de las medidas utilizadas en la interpretación de las medidas y consiguiente búsqueda de recomendaciones a posteriori. Así, si la capa de recolección no verificare la huella de las transacciones desde el adaptador, las medidas no se consideran en el análisis de los indicadores y su toma de decisiones asociadas.

Se implementó el registro unificado de nodos empleando tecnología Blockchain a los efectos de dotar de independencia a los nodos involucrados en la recolección de datos respecto de las capas de la arquitectura en la nube (Ver sección 6.4). De este modo, cualquier adaptador o pasarela puede acceder a la base de datos distribuida con información de los nodos a los efectos de realizar transmisiones indirectas de medidas. Cada nodo actualiza su registro y se consolida la base de datos por consenso a partir del nodo de referencia calculado por la función de puntuación respectiva. Se provee una implementación de referencia con su respectiva simulación para estimar patrones de aplicabilidad.

Finalmente, se definió un escenario de uso a los efectos de mostrar la aplicación de la tesis a una problemática regional como lo es el material particulado en la provincia de La Pampa. A su vez, se esquematizó la situación en la cual el mismo proyecto (idéntica estructura) debe contrastarse funcionalmente con el comportamiento de las métricas a partir de los datos de los sensores y cómo ello se vincula con el cálculo de la similitud comportamental. De este modo, se graficó el modo en que se obtenía la matriz de similitud de proyectos y cómo dinámicamente se obtenía a partir de allí el orden de acceso para la memoria organizacional, sin que se haya consultado esta de ningún modo previamente. De este modo, ante la ausencia de conocimiento en un proyecto, esto permitiría mejorar la precisión de la búsqueda de recomendaciones asociadas dado que se focaliza directamente sobre proyectos que no solo se parecen, sino que se comportan en forma similar. Además, la lista ordenada de proyectos similares cambia instante a instante en forma incremental con el arribo de nuevos datos y sin implicar un exceso en la carga de procesamiento.

De este modo, se ha podido lograr una estrategia de recomendación en memoria, basado en entidades bajo monitoreo semánticamente similares, para mejorar la precisión y reutilización de conocimiento y/o experiencia previa ante situaciones nuevas y-o no tipificadas.

Trabajos Futuros

Sin embargo, quedan abiertos desafíos referidos a la gestión de memoria a nivel de adaptador de medición como de la capa de recolección de datos y la memoria organizacional. Por un lado, una estrategia de compresión de datos en memoria permitiría incrementar la capacidad de equipos con recursos limitados, articulándose con detectores de cambio, barreras temporales, y técnicas de descarte selectivo. Ello permitiría una mejora en la gestión de recurso a la vez que garantizaría un sustento estadístico dado en los datos a transmitir. Por otro lado, la compresión de datos en memoria es una alternativa viable para complementar la operación de consolidación en la cadena de bloques a los efectos de optimizar el registro histórico de cambios. Adicionalmente, la compresión en memoria permitiría articular los detectores de cambio en los datos respecto de la memoria organizacional, para definir una estrategia de actualización de caché en el adaptador de medición con probables cursos de acción asociados.

Por otro lado, un mapeo automático basado en BriefPD entre atributos y propiedades de contexto respecto de los sensores permanece pendiente. Ello permitiría incluso implementar opción automática de alternativas ante múltiples sensores para un mismo atributo.

Una homogenización a nivel de librerías, utilidades, y contenedores a nivel de API, microservicios y arquitectura de procesamiento será abordado. Ello permitirá replicar fácilmente la arquitectura completa mediante contenedores y Kubernetes de modo de facilitar su uso tanto nivel local como de clúster. Adicionalmente, facilitará la aplicación y despliegue de la arquitectura para su uso en diferentes proyectos que requieran monitoreo en tiempo real. A su vez, promoverá la creación de nuevos componentes para extender y-o complementar la funcionalidad de la arquitectura a lo largo de su cadena de procesamiento.

Anexo

a.1 Muestra de Datos

Tabla 51 Datos de la Estación de Monitoreo de La Cuesta, Toay (LC-Toay) – Abril 13 de 2022

Hora	Mat. Part. (ug/m3)	Temp. Amb. (ºC)	Hum. Amb (%).	Pres. Amb (hPa).	Lluv. (mm)	V. Viento (km/h)	O. Viento	Hum. Suelo	Temp. Suelo
00:00	1.071	13.30	100.00	1,011.90	-	-	225.00	22.73	15.60
01:00	0.7455	13.60	100.00	1,012.20	-	-	315.00	22.73	15.60
02:00	0.5985	13.80	100.00	1,011.30	-	-	315.00	22.73	15.60
03:00	0.5775	13.60	100.00	1,011.30	-	-	315.00	22.73	15.60
04:00	0.3045	12.90	100.00	1,011.50	0.20	-	315.00	22.73	15.60
05:00	0.42	13.10	100.00	1,011.80	-	-	-	22.73	15.60
06:00	0.6615	13.30	100.00	1,011.90	-	1.60	-	22.73	15.00
07:00	0.945	13.10	100.00	1,012.70	-	1.60	-	22.73	15.00
08:00	24.465	13.70	100.00	1,013.10	-	6.40	-	22.73	15.00
09:00	33.18	14.10	100.00	1,013.50	-	4.80	-	22.73	15.00
10:00	17.325	14.70	100.00	1,014.00	-	8.00	-	22.73	15.60
11:00	9.324	16.40	79.00	1,013.90	-	4.80	315.00	22.73	15.60
12:00	4.7775	18.20	59.00	1,013.80	-	6.40	315.00	22.73	16.10
13:00	2.877	19.60	59.00	1,013.10	-	4.80	90.00	22.73	16.70
14:00	1.7535	20.60	50.00	1,012.50	-	1.60	90.00	22.73	17.20
15:00	1.155	21.40	46.00	1,012.20	-	1.60	90.00	22.73	17.80
16:00	1.869	21.90	42.00	1,012.20	-	1.60	45.00	22.73	17.80
17:00	1.68	20.80	45.00	1,012.20	-	1.60	-	22.73	17.80
18:00	8.736	19.30	60.00	1,012.30	-	-	315.00	22.73	17.80
19:00	8.7675	15.80	71.00	1,012.60	-	-	315.00	22.73	17.20
20:00	4.914	13.70	73.00	1,012.80	-	-	-	22.73	16.70
21:00	3.3075	13.30	78.00	1,013.10	-	-	315.00	22.73	16.70
22:00	3.0975	11.60	84.00	1,013.20	-	-	270.00	22.73	16.10
23:00	1.554	10.80	87.00	1,013.30	-	-	270.00	22.73	15.60

Tabla 52 Datos de la Estación de Monitoreo del Campus - UNLPam (Santa Rosa) – Abril 13 de 2022

Hora	Mat. Part. (ug/m3)	Temp. Amb. (ºC)	Hum. Amb (%).	Pres. Amb (hPa).	Lluv. (mm)	V. Viento (km/h)	O. Viento	Hum. Suelo	Temp. Suelo
00:00	1.0455	17.40	77.00	1,000.90	-	1.60	225.00	22.73	19.40
01:00	0.72775	16.40	80.00	1,000.40	-	-	225.00	22.73	19.40
02:00	0.58425	17.40	71.00	1,000.40	-	6.40	225.00	22.73	18.90
03:00	0.56375	17.30	75.00	1,000.70	-	1.60	225.00	22.73	18.90
04:00	0.29725	15.80	81.00	1,001.20	-	-	225.00	22.73	18.90

05:00	0.41	13.90	83.00	1,001.00	-	-	225.00	22.73	18.30
06:00	0.64575	13.80	86.00	1,001.70	-	-	225.00	22.73	18.30
07:00	0.9225	13.60	86.00	1,002.50	-	-	225.00	22.73	18.30
08:00	23.8825	12.90	88.00	1,003.20	-	-	-	22.73	17.80
09:00	32.39	14.70	62.00	1,004.30	-	1.60	90.00	22.73	17.80
10:00	16.9125	18.50	50.00	1,005.00	-	6.40	-	22.73	17.80
11:00	9.102	19.70	35.00	1,005.40	-	8.00	90.00	22.73	18.30
12:00	4.66375	20.20	34.00	1,004.50	-	8.00	90.00	22.73	18.30
13:00	2.8085	21.10	27.00	1,004.10	-	8.00	90.00	22.73	18.30
14:00	1.71175	23.20	24.00	1,003.70	-	8.00	90.00	22.73	18.90
15:00	1.1275	23.00	27.00	1,002.50	-	11.30	90.00	22.73	18.90
16:00	1.8245	23.20	29.00	1,002.40	-	11.30	90.00	22.73	19.40
17:00	1.64	22.80	33.00	1,001.80	-	6.40	90.00	22.73	19.40
18:00	8.528	21.70	43.00	1,001.60	-	3.20	90.00	22.73	19.40
19:00	8.55875	20.30	45.00	1,002.20	-	-	-	22.73	19.40
20:00	4.797	18.80	50.00	1,002.60	-	-	-	22.73	19.40
21:00	3.22875	16.80	58.00	1,002.10	-	-	-	22.73	18.90
22:00	3.02375	16.10	61.00	1,002.50	-	-	-	22.73	18.90
23:00	1.517	14.80	67.00	1,003.40	-	-	-	22.73	18.30

Tabla 53 Datos de la Estación de Monitoreo de CCEEyNN - UNLPam (Santa Rosa) – Abril 13 de 2022

Hora	Mat. Part. (ug/m3)	Temp. Amb. (ºC)	Hum. Amb (%).	Pres. Amb (hPa).	Lluv. (mm)	V. Viento (km/h)	O. Viento	Hum. Suelo	Temp. Suelo
00:00	1.02	4.70	60.00	1,023.10	-	-	-	22.73	15.60
01:00	0.71	5.20	56.00	1,023.20	-	-	-	22.73	15.00
02:00	0.57	5.90	54.00	1,022.60	-	-	-	22.73	15.00
03:00	0.55	5.40	57.00	1,022.20	-	-	-	22.73	14.40
04:00	0.29	5.20	58.00	1,021.80	-	-	-	22.73	13.90
05:00	0.4	4.20	63.00	1,021.80	-	-	-	22.73	13.30
06:00	0.63	3.80	66.00	1,022.30	-	-	-	22.73	13.30
07:00	0.9	3.10	72.00	1,022.20	-	-	315.00	22.73	12.80
08:00	23.3	5.20	60.00	1,021.40	-	-	315.00	22.73	12.80
09:00	31.6	9.70	49.00	1,021.60	-	4.80	90.00	22.73	12.80
10:00	16.5	12.80	39.00	1,022.40	-	6.40	90.00	22.73	13.30
11:00	8.88	16.20	28.00	1,022.30	-	9.70	90.00	22.73	13.90
12:00	4.55	17.90	26.00	1,021.30	-	11.30	90.00	22.73	15.00
13:00	2.74	18.80	25.00	1,020.10	-	12.90	90.00	22.73	16.10
14:00	1.67	19.90	24.00	1,018.80	-	12.90	90.00	22.73	16.70
15:00	1.1	20.40	26.00	1,018.00	-	11.30	90.00	22.73	17.20
16:00	1.78	20.60	26.00	1,017.80	-	11.30	90.00	22.73	17.20

17:00	1.6	20.10	28.00	1,017.00	-	11.30	90.00	22.73	17.20
18:00	8.32	18.80	34.00	1,017.20	-	9.70	90.00	22.73	17.20
19:00	8.35	16.40	41.00	1,018.10	-	3.20	90.00	22.73	16.70
20:00	4.68	14.90	43.00	1,018.10	-	4.80	90.00	22.73	16.70
21:00	3.15	14.20	47.00	1,018.60	-	6.40	90.00	22.73	16.10
22:00	2.95	12.90	53.00	1,018.90	-	4.80	90.00	22.73	16.10
23:00	1.48	11.90	58.00	1,019.30	-	4.80	90.00	22.73	15.60

Tabla 54 Datos de la Estación de Monitoreo de General Acha – Abril 13 de 2022

Hora	Mat. Part. (ug/m3)	Temp. Amb. (ºC)	Hum. Amb (%).	Pres. Amb (hPa).	Lluv. (mm)	V. Viento (km/h)	O. Viento	Hum. Suelo	Temp. Suelo
00:00	1.34895	13.60	90.00	1,013.20	-	-	-	22.73	15.60
01:00	0.938975	12.60	94.00	1,013.30	-	-	-	22.73	15.00
02:00	0.753825	12.40	89.00	1,013.10	-	-	225.00	22.73	15.00
03:00	0.727375	12.20	87.00	1,013.00	-	-	-	22.73	15.00
04:00	0.383525	11.50	90.00	1,012.70	-	-	-	22.73	15.00
05:00	0.529	10.40	92.00	1,012.50	-	-	-	22.73	14.40
06:00	0.833175	9.40	97.00	1,012.80	-	-	-	22.73	14.40
07:00	1.19025	8.80	100.00	1,013.40	-	-	-	22.73	13.90
08:00	30.81425	8.50	100.00	1,013.70	-	-	-	22.73	13.90
09:00	41.791	9.90	100.00	1,014.20	-	-	225.00	22.73	13.90
10:00	21.82125	13.80	86.00	1,014.40	-	-	225.00	22.73	14.40
11:00	11.7438	17.70	66.00	1,014.40	-	-	225.00	22.73	15.00
12:00	6.017375	21.40	60.00	1,014.10	-	1.60	225.00	22.73	15.60
13:00	3.62365	23.90	41.00	1,013.40	-	1.60	225.00	22.73	16.70
14:00	2.208575	25.60	43.00	1,012.60	-	3.20	225.00	22.73	17.20
15:00	1.45475	26.00	41.00	1,012.00	-	4.80	225.00	22.73	17.20
16:00	2.35405	26.10	39.00	1,011.60	-	8.00	225.00	22.73	17.80
17:00	2.116	25.80	44.00	1,011.50	-	6.40	225.00	22.73	17.80
18:00	11.0032	25.00	50.00	1,011.60	-	4.80	225.00	22.73	17.20
19:00	11.042875	22.60	64.00	1,011.40	-	-	225.00	22.73	17.20
20:00	6.1893	19.00	70.00	1,011.80	-	-	-	22.73	16.70
21:00	4.165875	17.40	74.00	1,011.90	-	-	-	22.73	16.70
22:00	3.901375	16.40	77.00	1,012.00	-	-	225.00	22.73	16.10
23:00	1.9573	15.10	85.00	1,011.80	-	-	225.00	22.73	16.10

Bibliografía

[1] M. Diván, *Information Technology Fundamentals for the Economics Sciences*. Santa Rosa - La Pampa: Publishing House of the National University of La Pampa, 2012.

[2] L. Olsina, F. Papa, and H. Molina, "How to Measure and Evaluate Web Applications in a Consistent Way," 2008. doi: 10.1007/978-1-84628-923-1_13.

[3] M. G. Mendonça and V. R. Basili, "Validation of an approach for improving existing measurement frameworks," *IEEE Transactions on Software Engineering*, 2000, doi: 10.1109/32.852739.

[4] M. Diván and M. Sánchez Reynoso, "Towards the monitoring of the pension processes in the Social Security Institute." Argentinian Conference on Computer Science - Electronic Government Simposium, La Plata, 2012.

[5] S. Ferdoush and X. Li, "Wireless Sensor Netwoek System Design using Raspberry Pi and Arduino for Environmental Monitoring Applications," *Procedia Computer Science* , vol. 34, pp. 103–110, 2014.

[6] A. Iqbal *et al.*, "Interoperable Internet-of-Things platform for smart home system using Web-of-Objects and cloud," *Sustainable Cities and Society*, vol. 38, 2018, doi: 10.1016/j.scs.2018.01.044.

[7] H. Molina and L. Olsina, "Towards the support of contextual information to a measurement and evaluation framework," 2007. doi: 10.1109/QUATIC.2007.31.

[8] P. Becker, "Process View of the Quality Measurement and Evaluation Integrated Strategies, PhD Thesis. ," La Plata, 2014.

[9] M. Diván and M. de Los Ángeles Martín, "Towards a consistent measurement stream processing from heterogeneous data sources," *International Journal of Electrical and Computer Engineering*, vol. 7, no. 6, 2017, doi: 10.11591/ijece.v7i6.pp3164-3175.

[10] N. Chaudhry, K. Shaw, and M. Abdelguerfi, "Stream Data Management," *New York: Springer-Verlag*, 2005.

[11] G. C. Smith, J. O. Hallstrom, S. Esswein, G. W. Eidson, and C. Post, "Managing metadata in heterogeneous sensor networks," 2014. doi: 10.1145/2638404.2638477.

[12] M. J. Divan, "Processing architecture based on measurement metadata," 2016. doi: 10.1109/ICRITO.2016.7784912.

[13] M. de Los Angeles Martin and M. J. Divan, "Case based organizational memory for processing architecture based on measurement metadata," 2016. doi: 10.1109/ICRITO.2016.7784954.

[14] D. E. Boubiche, "Secure and Efficient Big Data Gathering in Heterogeneous Wireless Sensor Networks," 2016. doi: 10.1145/2896387.2900334.

[15] L. Sun, M. J. Franklin, J. Wang, and E. Wil, “Skipping-oriented partitioning for columnar layouts,” in *Proceedings of the VLDB Endowment*, 2016, vol. 10, no. 4. doi: 10.14778/3025111.3025123.

[16] M. J. Divan and M. L. S. Reynoso, “Behavioural similarity analysis for supporting the recommendation in PAbMM,” in *2017 International Conference on Infocom Technologies and Unmanned Systems: Trends and Future Directions, ICTUS 2017*, Feb. 2018, vol. 2018-January, pp. 133–139. doi: 10.1109/ICTUS.2017.8285992.

[17] K. Schwaber and J. Sutherland, “The Scrum Guide: The Definitive The Rules of the Game,” *Scrum.Org and ScrumInc*, no. November, 2017.

[18] M. J. Divan and M. L. S. Reynoso, “Behavioural similarity analysis for supporting the recommendation in PAbMM,” in *2017 International Conference on Infocom Technologies and Unmanned Systems: Trends and Future Directions, ICTUS 2017*, 2018, vol. 2018-Janua. doi: 10.1109/ICTUS.2017.8285992.

[19] P. Runeson and M. Höst, “Guidelines for conducting and reporting case study research in software engineering,” *Empirical Software Engineering*, 2009, doi: 10.1007/s10664-008-9102-8.

[20] J. M. Verner, J. Sampson, V. Tosic, N. A. Abu Bakar, and B. A. Kitchenham, “Guidelines for industrially-based multiple case studies in software engineering,” 2009. doi: 10.1109/RCIS.2009.5089295.

[21] B. Kitchenham and S. Charters, “Guidelines for performing Systematic Literature reviews in Software Engineering Version 2.3,” *Engineering*, 2007, doi: 10.1145/1134285.1134500.

[22] B. A. Kitchenham, D. Budgen, and O. Pearl Brereton, “Using mapping studies as the basis for further research - A participant-observer case study,” *Information and Software Technology*, 2011, doi: 10.1016/j.infsof.2010.12.011.

[23] K. Petersen, S. Vakkalanka, and L. Kuzniarz, “Guidelines for conducting systematic mapping studies in software engineering: An update,” 2015. doi: 10.1016/j.infsof.2015.03.007.

[24] F. Ricci, L. Rokach, and B. Shapira, “Recommender Systems Handbook,” *P.B. Kantor (Eds) Springer*, p. 842, 2011.

[25] P. Resnick, “ Recommender Systems,” *H.R. Varian (Guest Eds.) Comm. of the ACM*, pp. 56–89, 1997.

[26] L. Chavarría Báez, “Los Sistemas de Recomendación en la Toma de Decisiones”.

[27] L.-Á. C., Q.-F. P.M., P.-C. P., O.-R. C.A., M.-C. P., and M.-H. J., “Literature review on information filtering methods in recommendation systems,” 2021. doi: 10.1109/ENC53357.2021.9534807.

[28] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen, “Collaborative filtering recommender systems,” in *The Adaptive Web: Methods and Strategies of Web Personalization*, 2007, pp. 291–324.

[29] M. F. and K. W., “Proposed model to intelligent recommendation system based on markov chains and grouping of genres,” in *Procedia Computer Science*, 2020, vol. 176, pp. 868 – 877. doi: 10.1016/j.procs.2020.09.082.

[30] G. Linden, B. Smith, and J. York, “ Amazon.com recommendations: item-to-item collaborative filtering. ,” *Internet Computing, IEEE*, vol. 7(1), pp. 76–80, 2003.

[31] Rui Xu and II Wunsch, “Survey of clustering algorithms. ,” *IEEE Transanctions on Neural Network*, vol. 16(3), pp. 645–678, 2005.

[32] S. Kumar, P. Tiwari, and M. Zymbler, “Internet of Things is a revolutionary approach for future technology enhancement: a review,” *Journal of Big Data*, vol. 6, no. 1, p. 111, 2019, doi: 10.1186/s40537-019-0268-2.

[33] A. M. S. S. Saleh I, *Challenges of the Internet of Things: Technique, Use, Ethics.* USA: John Wiley & Sons, Inc; , 2018.

[34] R. Mehta, J. Sahni, and K. Khanna, “Internet of Things: Vision, Applications and Challenges,” *Procedia Computer Science*, vol. 132, pp. 1263–1269, 2018, doi: https://doi.org/10.1016/j.procs.2018.05.042.

[35] D. Kavitha and S. Ravikumar, “IOT and context-aware learning-based optimal neural network model for real-time health monitoring,” *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 1, 2021, doi: 10.1002/ett.4132.

[36] O. Friha, M. A. Ferrag, L. Shu, L. Maglaras, and X. Wang, “Internet of Things for the Future of Smart Agriculture: A Comprehensive Survey of Emerging Technologies,” *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 4. 2021. doi: 10.1109/JAS.2021.1003925.

[37] N. Islam, M. M. Rashid, F. Pasandideh, B. Ray, S. Moore, and R. Kadel, “A review of applications and communication technologies for internet of things (Iot) and unmanned aerial vehicle (uav) based sustainable smart farming,” *Sustainability (Switzerland)*, vol. 13, no. 4, 2021, doi: 10.3390/su13041821.

[38] M. J. Divan, M. L. Sanchez-Reynoso, J. E. Panebianco, and M. J. Mendez, “IoT-Based Approaches for Monitoring the Particulate Matter and Its Impact on Health,” *IEEE Internet of Things Journal*, vol. 8, no. 15, pp. 11983–12003, Aug. 2021, doi: 10.1109/JIOT.2021.3068898.

[39] M. L. Diván Mario José and Sánchez Reynoso, “An Architecture for the Real-Time Data Stream Monitoring in IoT,” in *Multimedia Big Data Computing for IoT Applications: Concepts, Paradigms and Solutions*, S. and K. N. Tanwar Sudeep and Tyagi, Ed. Singapore: Springer Singapore, 2020, pp. 59–100. doi: 10.1007/978-981-13-8759-3_3.

[40] X. Wang *et al.*, “A Fog-Based Recommender System,” *IEEE Internet of Things Journal*, vol. 7, no. 2, 2020, doi: 10.1109/JIOT.2019.2949029.

[41] J. Calvillo-Arbizu, I. Román-Martínez, and J. Reina-Tosina, “Internet of things in health: Requirements, issues, and gaps,” *Computer Methods and Programs in Biomedicine*, vol. 208, 2021, doi: 10.1016/j.cmpb.2021.106231.

[42] A. Gupta and A. Al-Anbuky, “Iot-based patient movement monitoring: The post-operative hip fracture rehabilitation model,” *Future Internet*, vol. 13, no. 8, 2021, doi: 10.3390/fi13080195.

[43] S. Neelakandan, M. A. Berlin, S. Tripathi, V. B. Devi, I. Bhardwaj, and N. Arulkumar, “IoT-based traffic prediction and traffic signal control system for smart city,” *Soft Computing*, vol. 25, no. 18, pp. 12241–12248, 2021, doi: 10.1007/s00500-021-05896-x.

[44] J. Bae, G. Kim, and S. J. Lee, “Real-time prediction of nuclear power plant parameter trends following operator actions,” *Expert Systems with Applications*, vol. 186, 2021, doi: 10.1016/j.eswa.2021.115848.

[45] M. H. Sulaiman, S. Sulaiman, and A. Saparon, “IoT for wheel alignment monitoring system,” *International Journal of Electrical and Computer Engineering*, vol. 11, no. 5, pp. 3809–3817, 2021, doi: 10.11591/ijece.v11i5.pp3809-3817.

[46] O. Yakubu and E. Wereko, “Internet of things based vital signs monitoring system: A prototype validity test,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 2, pp. 962–972, 2021, doi: 10.11591/ijeecs.v23.i2.pp962-972.

[47] Y. bin Lin *et al.*, “The artificial intelligence of things sensing system of real-time bridge scour monitoring for early warning during floods,” *Sensors*, vol. 21, no. 14, Jul. 2021, doi: 10.3390/s21144942.

[48] F. J. García-Peñalvo, “Cómo hacer una Systematic Literature Review (SLR) ,” *Zenodo, 2021, [Online]*. 2021.

[49] M. L. Sánchez-Reynoso, M. J. Diván, S. Gonnet, and M. Méndez, “Real-Time Recommenders applied to IoT based Systems_v2.5,” 2021.

[50] P. Kaur, R. Kumar, and M. Kumar, “A healthcare monitoring system using random forest and internet of things (IoT),” *Multimedia Tools and Applications*, vol. 78, no. 14, pp. 19905–19916, Jul. 2019, doi: 10.1007/s11042-019-7327-8.

[51] H. Jeong, B. Park, M. Park, K. B. Kim, and K. Choi, “Big data and rule-based recommendation system in Internet of Things,” *Cluster Computing*, vol. 22, pp. 1837–1846, Jan. 2019, doi: 10.1007/s10586-017-1078-y.

[52] R. Wang, Y. Liu, P. Zhang, X. Li, and X. Kang, “Edge and cloud collaborative entity recommendation method towards the IoT search,” *Sensors (Switzerland)*, vol. 20, no. 7, Apr. 2020, doi: 10.3390/s20071918.

[53] R. O. J. Betancourt *et al.*, “Iot-based electricity bill for domestic applications,” *Sensors (Switzerland)*, vol. 20, no. 21, pp. 1–16, Nov. 2020, doi: 10.3390/s20216178.

[54] H. Tan and Y. Li, “News Information Platform Optimization Based on the Internet of Things,” *Wireless Communications and Mobile Computing*, vol. 2021, 2021, doi: 10.1155/2021/9403874.

[55] P. Mahajan and P. D. Kaur, “Three-tier IoT-edge-cloud (3T-IEC) architectural paradigm for real-time event recommendation in event-based social networks,”

Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 1, pp. 1363–1386, Jan. 2021, doi: 10.1007/s12652-020-02202-9.

[56] Y. Yang, J. Xu, Z. Xu, P. Zhou, and T. Qiu, "Quantile context-aware social IoT service big data recommendation with D2D communication," *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 5533–5548, 2020, doi: 10.1109/JIOT.2020.2980046.

[57] Y. Saleem *et al.*, "IoTRec: The IoT Recommender for Smart Parking System," *IEEE Transactions on Emerging Topics in Computing*, 2020, doi: 10.1109/TETC.2020.3014722.

[58] K. S. Umadevi, P. Balakrishnan, and G. Kousalya, "Intrusion detection system using timed automata for cyber physical systems," *Journal of Intelligent and Fuzzy Systems*, vol. 36, no. 5, pp. 4005–4015, 2019, doi: 10.3233/JIFS-169961.

[59] B. O'Flynn, F. Regan, A. Lawlor, J. Wallace, J. Torres, and C. O'Mathuna, "Experiences and recommendations in deploying a real-time, water quality monitoring system," *Measurement Science and Technology*, vol. 21, no. 12, 2010, doi: 10.1088/0957-0233/21/12/124004.

[60] S. T. John, A. Mohan, M. S. Philip, P. Sarkar, and R. Davis, "An IoT device for striking of vertical concrete formwork," *Engineering, Construction and Architectural Management*, vol. ahead-of-print, no. ahead-of-print, Jan. 2021, doi: 10.1108/ECAM-10-2020-0859.

[61] R. Jolak *et al.*, "CONSERVE: A framework for the selection of techniques for monitoring containers security," *Journal of Systems and Software*, vol. 186, 2022, doi: 10.1016/j.jss.2021.111158.

[62] P. A. Dreyfus, F. Psarommatis, G. May, and D. Kiritsis, "Virtual metrology as an approach for product quality estimation in Industry 4.0: a systematic review and integrative conceptual framework," *International Journal of Production Research*, vol. 60, no. 2, pp. 742–765, 2022, doi: 10.1080/00207543.2021.1976433.

[63] M. Yazdani, D. Pamucar, P. Chatterjee, and S. Chakraborty, "Development of a decision support framework for sustainable freight transport system evaluation using rough numbers," *International Journal of Production Research*, vol. 58, no. 14, pp. 4325 – 4351, 2020, doi: 10.1080/00207543.2019.1651945.

[64] M. Diván, "Enfoque Integrado de Procesamiento de Flujos de Datos centrado en Metadatos de Mediciones. ," PhD Thesis, Facultad de Informática, Universidad Nacional de La Plata, La Plata, Argentina , 2011.

[65] L. Olsina and M. de Los Angeles Martín, "Ontology for software metrics and indicators: Building process and decisions taken," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2004, vol. 3140. doi: 10.1007/978-3-540-27834-4_23.

[66] M. Diván, L. Olsina, and S. Gordillo, "Strategy for Data Stream Processing Based on Measurement Metadata: An Outpatient Monitoring Scenario," *Journal of Software Engineering and Applications*, vol. 04, no. 12, pp. 653–665, 2011, doi: 10.4236/jsea.2011.412077.

[67] P. Becker, H. Molina, and L. Olsina, “Measurement and evaluation as a quality driver,” *Ingénierie des systèmes d’information*, vol. 15, no. 6, 2010, doi: 10.3166/isi.15.6.33-62.

[68] M. de Los Ángeles Martín and M. J. Diván, “Applications of case based organizational memory supported by the PAbMM architecture,” *Advances in Science, Technology and Engineering Systems*, vol. 2, no. 3, pp. 12–23, 2017, doi: 10.25046/aj020303.

[69] J. Wang and Y. Dong, “Measurement of text similarity: A survey,” *Information (Switzerland)*, vol. 11, no. 9. MDPI AG, pp. 1–17, Sep. 01, 2020. doi: 10.3390/info11090421.

[70] P. Becker, F. Papa, and L. Olsina, “Enhancing the conceptual framework capability for a measurement and evaluation strategy,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 8295 LNCS, pp. 104–116. doi: 10.1007/978-3-319-04244-2_11.

[71] N. Kanayama, M. Hara, and K. Kimura, “Virtual reality alters cortical oscillations related to visuo-tactile integration during rubber hand illusion,” *Scientific Reports*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-020-80807-y.

[72] R. Navigli and F. Martelli, “An overview of word and sense similarity,” *Natural Language Engineering*, vol. 25, no. 6. Cambridge University Press, pp. 693–714, Nov. 01, 2019. doi: 10.1017/S1351324919000305.

[73] Adhikari A, Dutta B, Dutta, and Mondal D, “Semantic similarity measurement: An intrinsic information content model. ,” *Int. J. Metadata, Semant. Ontol.,* , vol. 14(3), pp. 218–233, 2020.

[74] N. A. Omar, S. Kasim, M. A. Hasibuan, and M. F. M. Fudzee, “Hyb-tvx: A hybrid semantic similarity feature-based measurement for multiple ontologies,” *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 1.3 S1, pp. 176–180, 2019, doi: 10.30534/ijatcse/2019/3581.32019.

[75] F. Z. Smaili, X. Gao, and R. Hoehndorf, “OPA2Vec: Combining formal and informal content of biomedical ontologies to improve similarity-based prediction,” *Bioinformatics*, vol. 35, no. 12, pp. 2133–2140, Jun. 2019, doi: 10.1093/bioinformatics/bty933.

[76] Chung K., Yoo H., and Choe D., “Ambient context-based modeling for health risk assessment using deep neural network. J. Ambient Intell. Humaniz. Comput., 2020, 11(4): 1387–1395.,” *J. Ambient Intell. Humaniz. Comput.* , vol. 11(4), pp. 1387–1395, 2020.

[77] B. Alhijawi and Y. Kilani, “A collaborative filtering recommender system using genetic algorithm,” *Information Processing and Management*, vol. 57, no. 6, Nov. 2020, doi: 10.1016/j.ipm.2020.102310.

[78] Hong B. and Yu M., “ A collaborative filtering algorithm based on correlation coefficient. ,” *Neural Comput.* , vol. 31(12), pp. 8317–8326, Apr. 2019.

[79] A. R. C. Maita *et al.*, “A systematic mapping study of process mining,” *Enterprise Information Systems*, vol. 12, no. 5. 2018. doi: 10.1080/17517575.2017.1402371.

[80] A. Idri, I. Abnane, and A. Abran, “Systematic mapping study of missing values techniques in software engineering data,” 2015. doi: 10.1109/SNPD.2015.7176280.

[81] V. H. S. Durelli *et al.*, “Machine learning applied to software testing: A systematic mapping study,” *IEEE Transactions on Reliability*, vol. 68, no. 3, 2019, doi: 10.1109/TR.2019.2892517.

[82] M. L. Sanchez-Reynoso and M. J. Divan, “A systematic literature mapping on the similar semantically entities in measurement projects,” 2019. doi: 10.1109/ICVRV47840.2019.00033.

[83] Y. Ma, L. Liu, K. Lu, B. Jin, and X. Liu, “A graph derivation based approach for measuring and comparing structural semantics of ontologies,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 5, pp. 1039–1052, 2014, doi: 10.1109/TKDE.2013.120.

[84] G. Glavaš, M. Franco-Salvador, S. P. Ponzetto, and P. Rosso, “A resource-light method for cross-lingual semantic textual similarity,” *Knowledge-Based Systems*, vol. 143, pp. 1–9, 2018, doi: 10.1016/j.knosys.2017.11.041.

[85] I. Traverso-Ribón, “Exploiting semantics from ontologies to enhance accuracy of similarity measures,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9088, pp. 795–805, 2015, doi: 10.1007/978-3-319-18818-8_52.

[86] B. Hajian and T. White, “Measuring semantic similarity using a multi-tree model,” in *CEUR Workshop Proceedings*, 2010, vol. 756, pp. 7–14. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84891940494&partnerID=40&md5=40758d5a7f7f0a878aa27ee0ab3c7953

[87] F. Lehner and R. K. Maier, “How Can Organizational Memory Theories Contribute to Organizational Memory Systems?,” *Information Systems Frontiers*, vol. 2, no. 3–4, 2000, doi: 10.1023/A:1026516627735.

[88] M. S. Ackerman and C. Halverson, “Organizational Memory as Objects, Processes, and Trajectories: An Examination of Organizational Memory in Use,” *Computer Supported Cooperative Work (CSCW)*, vol. 13, no. 2, 2004, doi: 10.1023/b:cosu.0000045805.77534.2a.

[89] M. S. Ackerman, “Augmenting organizational memory: A field study of Answer Garden,” *ACM Transactions on Information Systems*, vol. 16, no. 3, 1998, doi: 10.1145/290159.290160.

[90] B. Ramesh, “Towards a meta-model for representing organizational memory,” in *Proceedings of the Hawaii International Conference on System Sciences*, 1997, vol. 2. doi: 10.1109/HICSS.1997.665561.

[91] A. Abecker, A. Bernardi, K. Hinkelmann, O. Kühn, and M. Sintek, “Toward a technology tor organizational memories,” *IEEE Intelligent Systems and Their Applications*, vol. 13, no. 3, 1998, doi: 10.1109/5254.683209.

[92] D. Nevo and Y. Wand, “Organizational memory information systems: A transactive memory approach,” *Decision Support Systems*, vol. 39, no. 4, 2005, doi: 10.1016/j.dss.2004.03.002.

[93] P. Jackson, “An Exploratory Survey of the Structure and Components of Organizational Memory,” in *Contributions to Management Science*, 2008. doi: 10.1007/978-3-7908-1958-8_7.

[94] E. Agbozo and K. Spassov, “Establishing efficient governance through data-driven e-government,” 2018. doi: 10.1145/3209415.3209419.

[95] M. J. Diván and M. L. S. Reynoso, “Real-Time Measurement and Evaluation as System Reliability Driver,” in *System Reliability Management*, 2018. doi: 10.1201/9781351117661-11.

[96] Zhou F, Wu B, Yang Y, Trajcevski G, Zhang K, and Zhong T, “ Vec2Link: Unifying Heterogeneous Data for Social Link Prediction ,” *27th ACM International Conference on Information and Knowledge Management, Torino, Italy.*, 2018.

[97] M. L. Sánchez Reynoso and M. Diván, “Improving the Real-Time Searching in the Organizational Memory,” in *Procedia Computer Science*, 2018, vol. 154. doi: 10.1016/j.procs.2019.06.043.

[98] J. C. dos Santos and M. L. P. Valentim, “Institutional memory and organizational memory: Faces of the same coin,” *Perspectivas em Ciencia da Informacao*, vol. 26, no. 3, 2021, doi: 10.1590/1981-5344/4315.

[99] M. de los Angeles Martin and L. Olsina, “Added value of ontologies for modeling an organizational memory,” in *Building Organizational Memories: Will You Know What You Knew?*, 2009. doi: 10.4018/978-1-59904-540-5.ch010.

[100] J. T. J. Penttinen and J. Borrill, “Measurements,” in *The LTE-Advanced Deployment Handbook*, John Wiley & Sons, Ltd, 2015, pp. 339–372. doi: 10.1002/9781118678879.ch11.

[101] B. Godin, “Outline for a history of science measurement,” *Science Technology and Human Values*. 2002. doi: 10.1177/016224390202700101.

[102] H. Guyon, J.-L. Kop, J. Juhel, and B. Falissard, “Measurement, ontology, and epistemology: Psychology needs pragmatism-realism,” *Theory & Psychology*, vol. 28, no. 2, pp. 149–171, Apr. 2018, doi: 10.1177/0959354318761606.

[103] C. C. Tramontini and K. U. Graziano, “Hypothermia control in elderly surgical patients in the intraoperative period: evaluation of two nursing interventions,” *Revista Latino-Americana de Enfermagem*, vol. 15, no. 4, pp. 626–631, Aug. 2007, doi: 10.1590/S0104-11692007000400016.

[104] K. Witkowski, “Internet of Things, Big Data, Industry 4.0 – Innovative Solutions in Logistics and Supply Chains Management,” *Procedia Engineering*, vol. 182, pp. 763–769, 2017, doi: 10.1016/j.proeng.2017.03.197.

[105] M. J. Diván and M. L. Sánchez-Reynoso, “Real-Time Measurement and Evaluation as System Reliability Driver,” in *System Reliability Management*, A. Anand and M. Ram, Eds. Boca Raton : Taylor & Francis, a CRC title, part of the Taylor &: CRC Press, 2018, pp. 161–188. doi: 10.1201/9781351117661-11.

[106] H. Molina and L. Olsina, “Towards the Support of Contextual Information to a Measurement and Evaluation Framework,” in *6th International Conference on the Quality of Information and Communications Technology (QUATIC 2007)*, Sep. 2007, pp. 154–166. doi: 10.1109/QUATIC.2007.21.

[107] L. Olsina, F. Papa, and H. Molina, “How to Measure and Evaluate Web Applications in a Consistent Way,” in *Web Engineering: Modelling and Implementing Web Applications*, G. Rossi, O. Pastor, D. Schwabe, and L. Olsina, Eds. London: Springer London, 2008, pp. 385–420. doi: 10.1007/978-1-84628-923-1_13.

[108] P. Becker, P. Lew, and L. Olsina, “Strategy to improve quality for software applications,” in *Proceeding of the 2nd workshop on Software engineering for sensor network applications - SESENA ’11*, 2011, p. 129. doi: 10.1145/1987875.1987897.

[109] M. J. Diván and M. L. Sánchez-Reynoso, “Managing the Data Meaning in the Data Stream Processing: A Systematic Literature Mapping,” 2020, pp. 31–46. doi: 10.1007/978-981-15-3357-0_3.

[110] M. J. Divan and M. L. S. Reynoso, “Incorporating Scenarios and States Definitions on Real-Time Entity Monitoring in PAbMM,” in *2019 XLV Latin American Computing Conference (CLEI)*, Sep. 2019, pp. 1–10. doi: 10.1109/CLEI47609.2019.235072.

[111] M. J. Diván and M. L. Sánchez-Reynoso, “Extending the Data Stream Processing Strategy to Scenario Analysis,” *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 1.4, pp. 1–8, Sep. 2019, doi: 10.30534/ijatcse/2019/0181.42019.

[112] M. J. Diván and M. L. Sánchez-Reynoso, “A Real-Time Entity Monitoring based on States and Scenarios,” *CLEI Electronic Journal*, vol. 23, no. 1, pp. 2:1--2:25, Apr. 2020, doi: 10.19153/cleiej.23.1.2.

[113] T. Munzner, *Visualization Analysis and Design*, 1st Editio. New York, USA: A K Peters/CRC Press, 2014. doi: 10.1201/b17511.

[114] N. Elmqvist and J. D. Fekete, “Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines,” *IEEE Transactions on Visualization and Computer Graphics*, 2010, doi: 10.1109/TVCG.2009.84.

[115] C. Kelleher and T. Wagener, “Ten guidelines for effective data visualization in scientific publications,” *Environmental Modelling and Software*, 2011, doi: 10.1016/j.envsoft.2010.12.006.

[116] M. J. Diván and M. Singh, “The Impact of the Measurement Process in Intelligent System of Data Gathering Strategies,” *Lecture Notes in Computer Science*

(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12615 LNCS, pp. 445–457, 2021, doi: 10.1007/978-3-030-68449-5_43.

[117] M. J. Diván, M. L. Sánchez-Reynoso, and S. M. Gonnet, "Measurement project interoperability for real-time data gathering systems," *Future Generation Computer Systems*, vol. 129, 2022, doi: 10.1016/j.future.2021.11.031.

[118] M. Divan and M. L. Sánchez-Reynoso, "Fostering the Interoperability of the Measurement and Evaluation Project Definitions in PAbMM," in *2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Aug. 2018, pp. 231–238. doi: 10.1109/ICRITO.2018.8748766.

[119] J. Kan and K. S. Kim, "MTFS: Merkle-Tree-Based File System," in *2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, May 2019, pp. 43–47. doi: 10.1109/BLOC.2019.8751389.

[120] C. Oham, R. A. Michelin, R. Jurdak, S. S. Kanhere, and S. Jha, "B-FERL: Blockchain based framework for securing smart vehicles," *Information Processing & Management*, vol. 58, no. 1, p. 102426, 2021, doi: https://doi.org/10.1016/j.ipm.2020.102426.

[121] J. Li, J. Wu, G. Jiang, and T. Srikanthan, "Blockchain-based public auditing for big data in cloud storage," *Information Processing & Management*, vol. 57, no. 6, p. 102382, 2020, doi: https://doi.org/10.1016/j.ipm.2020.102382.

[122] A. Reyna, C. Martín, J. Chen, E. Soler, and M. Díaz, "On blockchain and its integration with IoT. Challenges and opportunities," *Future Generation Computer Systems*, vol. 88, pp. 173–190, Nov. 2018, doi: 10.1016/j.future.2018.05.046.

[123] M. de los Ángeles, Martín and D. M. José, "Applications of Case Based Organizational Memory Supported by the PAbMM Architecture," *Advances in Science, Technology and Engineering Systems Journal*, vol. 2, no. 3, pp. 12–23, Apr. 2017, doi: 10.25046/aj020303.

[124] J. Wang and Y. Dong, "Measurement of text similarity: A survey," *Information (Switzerland)*, vol. 11, no. 9, pp. 1–17, 2020, doi: 10.3390/info11090421.

[125] S. Wang *et al.*, "Real-time forecasting and early warning of bacillary dysentery activity in four meteorological and geographic divisions in China," *Science of the Total Environment*, vol. 761, 2021, doi: 10.1016/j.scitotenv.2020.144093.

[126] P. Cerva, L. Mateju, J. Zdansky, R. Safarik, and J. Nouza, "Identification of related languages from spoken data: Moving from off-line to on-line scenario," *Computer Speech & Language*, vol. 68, p. 101180, Jul. 2021, doi: 10.1016/j.csl.2020.101180.

[127] L. H. Son, "Dealing with the new user cold-start problem in recommender systems: A comparative review," *Information Systems*, vol. 58, pp. 87–104, 2016, doi: 10.1016/j.is.2014.10.001.

[128] M. Diván and M. L. Sánchez-Reynoso, “Monitoreo de entidades en tiempo real basado en estados y escenarios,” *CLEI Electronic Journal*, vol. 23, no. 1, pp. 2:1--2:25, Apr. 2020, doi: https://doi.org/10.19153/cleiej.23.1.2.

[129] J. Han, M. Kamber, and J. Pei, “2 - Getting to Know Your Data,” in *Data Mining (Third Edition)*, Third Edit., J. Han, M. Kamber, and J. Pei, Eds. Boston: Morgan Kaufmann, 2012, pp. 39–82. doi: https://doi.org/10.1016/B978-0-12-381479-1.00002-2.

[130] A. Gragera and V. Suppakitpaisarn, “Relaxed triangle inequality ratio of the Sørensen–Dice and Tversky indexes,” *Theoretical Computer Science*, vol. 718, pp. 37–45, 2018, doi: https://doi.org/10.1016/j.tcs.2017.01.004.

[131] S. Kosub, “A note on the triangle inequality for the Jaccard distance,” *Pattern Recognition Letters*, vol. 120, pp. 36–38, 2019, doi: https://doi.org/10.1016/j.patrec.2018.12.007.

[132] T. J. Duff, D. M. Chong, and K. G. Tolhurst, “Indices for the evaluation of wildfire spread simulations using contemporaneous predictions and observations of burnt area,” *Environmental Modelling & Software*, vol. 83, pp. 276–285, 2016, doi: https://doi.org/10.1016/j.envsoft.2016.05.005.

[133] D. M. Levine, K. A. Szabat, and D. F. Stephan, *Business Statistics: A First Course*, 7th ed. Pearson, 2014.

[134] M. J. Diván and M. L. Sánchez-Reynoso, “A Metadata and Z Score-based Load-Shedding Technique in IoT-based Data Collection Systems,” *International Journal of Mathematical, Engineering and Management Sciences*, 2021, doi: 10.33889/IJMEMS.2021.6.1.023.

[135] M. K. Garba, T. M. W. Nye, and R. J. Boys, “Probabilistic Distances between Trees,” *Systematic Biology*, vol. 67, no. 2, pp. 320–327, 2018, doi: 10.1093/sysbio/syx080.

[136] Q. Chen and M. Huang, “Rough fuzzy model based feature discretization in intelligent data preprocess,” *Journal of Cloud Computing*, vol. 10, no. 1, p. 5, Dec. 2021, doi: 10.1186/s13677-020-00216-4.

[137] M. J. Diván, “Enfoque integrado de procesamiento de flujos de datos centrado en metadatos de mediciones,” Universidad Nacional de La Plata, 2011. doi: 10.35537/10915/4198.

[138] M. Diván and L. Olsina, “Vista de Proceso del Enfoque Integrado de Procesamiento de Flujos de Datos centrado en Metadatos de Mediciones,” in *Argentine Symposium on Software Engineering*, 2013, vol. 42.

[139] M. Diván and L. Olsina, “Process View for a Data Stream Processing Strategy based on Measurement Metadata,” *Electronic Journal of SADIO*, vol. 13, no. 1, pp. 16–31, 2014, [Online]. Available: https://publicaciones.sadio.org.ar/index.php/EJS/article/view/39

[140] L.-J. Chen *et al.*, “An Open Framework for Participatory PM2.5 Monitoring in Smart Cities,” *IEEE Access*, vol. 5, pp. 14441–14454, 2017, doi: 10.1109/ACCESS.2017.2723919.

[141] Z. Ciu, E. Damiani, and M. Leida, “Benefits of ontologies in real time data access,” 2007. doi: 10.1109/DEST.2007.372004.

[142] M. J. Diván and M. L. Sánchez Reynoso, “An Architecture for the Real-Time Data Stream Monitoring in IoT,” in *Multimedia Big Data Computing for IoT Applications*, vol. 163, S. Tanwar, S. Tyagi, and N. Kumar, Eds. Springer Nature Singapore, 2020, pp. 59–100. doi: 10.1007/978-981-13-8759-3_3.

[143] C. Steinmetz, A. Rettberg, F. G. C. Ribeiro, G. Schroeder, M. S. Soares, and C. E. Pereira, “Using Ontology and Standard Middleware for integrating IoT based in the Industry 4.0**This work was conducted during a scholarship supported by the International Cooperation Program PROBRAL CAPES/DAAD at the University of Muenster. Financed by CAPES Brazi,” *IFAC-PapersOnLine*, vol. 51, no. 10, pp. 169–174, 2018, doi: https://doi.org/10.1016/j.ifacol.2018.06.256.

[144] F. Karim, M.-E. Vidal, and S. Auer, “Compact representations for efficient storage of semantic sensor data,” *Journal of Intelligent Information Systems*, Jan. 2021, doi: 10.1007/s10844-020-00628-3.

[145] U. Majeed, L. U. Khan, I. Yaqoob, S. M. A. Kazmi, K. Salah, and C. S. Hong, “Blockchain for IoT-based smart cities: Recent advances, requirements, and future challenges,” *Journal of Network and Computer Applications*, vol. 181, p. 103007, 2021, doi: https://doi.org/10.1016/j.jnca.2021.103007.

[146] M. J. Diván and M. L. Sánchez-Reynoso, “Metadata-based measurements transmission verified by a Merkle Tree,” *Knowledge-Based Systems*, vol. 219, p. 106871, May 2021, doi: 10.1016/j.knosys.2021.106871.

[147] M. Diván, M. L. Sánchez Reynoso, and M. H. A. Wahab, “Dynamic Switching in the Measurements’ Collecting from Heterogeneous Data Sources,” *Journal of Physics: Conference Series*, vol. 1529, p. 022058, Apr. 2020, doi: 10.1088/1742-6596/1529/2/022058.

[148] R. R. Rhinehart, “Tutorial: process control through nonlinear modeling,” in *Proceedings of the 1997 American Control Conference (Cat. No.97CH36041)*, 1997, pp. 2011–2015 vol.3. doi: 10.1109/ACC.1997.611041.

[149] N. F. Zhang, “A Statistical Control Chart for Stationary Process Data,” *Technometrics*, vol. 40, no. 1, pp. 24–38, Feb. 1998, doi: 10.1080/00401706.1998.10485479.

[150] R. R. Rhinehart, “CUSUM type on-line filter,” *Process Control and Quality*, vol. 2, pp. 169–176, 1992.

[151] S. Cao and R. R. Rhinehart, “An efficient method for on-line identification of steady state,” *Journal of Process Control*, vol. 5, no. 6, pp. 363–374, Dec. 1995, doi: 10.1016/0959-1524(95)00009-F.

[152] S. Cao and R. Russell Rhinehart, “A self-tuning filter,” *Journal of Process Control*, vol. 7, no. 2, pp. 139–148, Jan. 1997, doi: 10.1016/S0959-1524(96)00024-8.

[153] J. S. Alford, B. M. Hrankowsky, and R. R. Rhinehart, “Data filtering in process automation systems,” *InTech*, vol. 65, no. 4, pp. 14–19, 2018.

[154] T.-C. Wu and R.-S. Sung, “An improved one-time digital signature scheme based on one-way function,” *Journal of Information Science and Engineering*, vol. 12, no. 3, pp. 387–395, 1996.

[155] T. Gavrilova, A. Alsufyev, and M. Gladkova, “Perceptual factors in knowledge map visual design,” in *ACM International Conference Proceeding Series*, 2015, vol. 21-22-Octo. doi: 10.1145/2809563.2809599.

[156] M. Diván and M. L. Sánchez-Reynoso, “Towards a Distributed Record of Measurement Adapters Powered by Blockchain Technology,” in *Transformations Through Blockchain Technology*, 2022. doi: 10.1007/978-3-030-93344-9_5.

[157] L. Li, Y. Jiang, and G. Liu, “Consensus with voting theory in blockchain environments,” in *Proceedings - 10th IEEE International Conference on Big Knowledge, ICBK 2019*, 2019, pp. 152–159. doi: 10.1109/ICBK.2019.00028.

[158] V. Martins *et al.*, “Chemical characterisation of particulate matter in urban transport modes,” *Journal of Environmental Sciences*, vol. 100, pp. 51–61, Feb. 2021, doi: 10.1016/j.jes.2020.07.008.

[159] C. D. Mathers and D. Loncar, “Projections of Global Mortality and Burden of Disease from 2002 to 2030,” *PLoS Medicine*, vol. 3, no. 11, p. e442, Nov. 2006, doi: 10.1371/journal.pmed.0030442.

[160] World Health Organization, “WHO | Projections of mortality and causes of death, 2016 to 2060,” 2016. www.who.int/healthinfo/global_burden_disease/projections/en/%0A

[161] A. Fiordelisi, P. Piscitelli, B. Trimarco, E. Coscioni, G. Iaccarino, and D. Sorriento, “The mechanisms of air pollution and particulate matter in cardiovascular diseases,” *Heart Failure Reviews*, vol. 22, no. 3, pp. 337–347, May 2017, doi: 10.1007/s10741-017-9606-7.

[162] M. S. Peixoto, M. F. de Oliveira Galvão, and S. R. Batistuzzo de Medeiros, “Cell death pathways of particulate matter toxicity,” *Chemosphere*, vol. 188, pp. 32–48, Dec. 2017, doi: 10.1016/j.chemosphere.2017.08.076.

[163] E.-J. Jo *et al.*, “Effects of particulate matter on respiratory disease and the impact of meteorological factors in Busan, Korea,” *Respiratory Medicine*, vol. 124, pp. 79–87, Mar. 2017, doi: 10.1016/j.rmed.2017.02.010.

[164] G. Bel and M. Holst, “Evaluation of the impact of Bus Rapid Transit on air pollution in Mexico City,” *Transport Policy*, vol. 63, pp. 209–220, Apr. 2018, doi: 10.1016/j.tranpol.2018.01.001.

[165] S. Abdullah *et al.*, “Air quality status during 2020 Malaysia Movement Control Order (MCO) due to 2019 novel coronavirus (2019-nCoV) pandemic,” *Science of The Total Environment*, vol. 729, p. 139022, Aug. 2020, doi: 10.1016/j.scitotenv.2020.139022.

[166] L. Megido, L. Negral, L. Castrillón, Y. Fernández-Nava, B. Suárez-Peña, and E. Marañón, “Impact of secondary inorganic aerosol and road traffic at a suburban

air quality monitoring station," *Journal of Environmental Management*, vol. 189, pp. 36–45, Mar. 2017, doi: 10.1016/j.jenvman.2016.12.032.

[167] P. P. Ray, D. Dash, and N. Kumar, "Sensors for internet of medical things: State-of-the-art, security and privacy issues, challenges and future directions," *Computer Communications*, vol. 160, pp. 111–131, Jul. 2020, doi: 10.1016/j.comcom.2020.05.029.

[168] M. F. Tuysuz and R. Trestian, "From serendipity to sustainable green IoT: Technical, industrial and political perspective," *Computer Networks*, vol. 182, p. 107469, Dec. 2020, doi: 10.1016/j.comnet.2020.107469.

[169] L. Minh Dang, K. Min, H. Wang, Md. Jalil Piran, C. Hee Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognition*, vol. 108, p. 107561, Dec. 2020, doi: 10.1016/j.patcog.2020.107561.

[170] Vikash, L. Mishra, and S. Varma, "Performance evaluation of real-time stream processing systems for Internet of Things applications," *Future Generation Computer Systems*, vol. 113, pp. 207–217, Dec. 2020, doi: 10.1016/j.future.2020.07.012.

[171] D. Zhang and S. S. Woo, "Real Time Localized Air Quality Monitoring and Prediction Through Mobile and Fixed IoT Sensing Network," *IEEE Access*, vol. 8, pp. 89584–89594, 2020, doi: 10.1109/ACCESS.2020.2993547.

[172] M. J. Divan, M. L. Sanchez-Reynoso, J. E. Panebianco, and M. J. Mendez, "IoT-Based Approaches for Monitoring the Particulate Matter and Its Impact on Health," *IEEE Internet of Things Journal*, vol. 8, no. 15, pp. 11983–12003, 2021, doi: 10.1109/JIOT.2021.3068898.

[173] T. Becnel *et al.*, "A Distributed Low-Cost Pollution Monitoring Platform," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10738–10748, Dec. 2019, doi: 10.1109/JIOT.2019.2941374.

[174] J. M. Michaud *et al.*, "Taxon-specific aerosolization of bacteria and viruses in an experimental ocean-atmosphere mesocosm," *Nature Communications*, vol. 9, no. 1, p. 2017, Dec. 2018, doi: 10.1038/s41467-018-04409-z.

[175] N. van Doremalen *et al.*, "Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1," *New England Journal of Medicine*, vol. 382, no. 16, pp. 1564–1567, Apr. 2020, doi: 10.1056/NEJMc2004973.